REVISITING THE CURRENT ISSUES IN MULTILEVEL STRUCTURAL EQUATION MODELING (MSEM): THE APPLICATION OF SAMPLING WEIGHTS AND THE TEST OF MEASUREMENT INVARIANCE IN MSEM

A Dissertation

by

LEINA ZHU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Oi-Man Kwok
Co-Chair of Committee,	Myeongsun Yoon
Committee Members,	Victor L. Willson
	Wen Luo
Head of Department,	Victor L. Willson

December 2015

Major Subject: Educational Psychology

ABSTRACT

Multilevel structural equation modeling (MSEM) has been widely used throughout the applied social and behavioral sciences. This dissertation revisited current issues in MSEM, including: the application of sampling weights and the test of measurement invariance.

The impact of using sampling weights on testing multilevel mediation effects in large-scale, complex survey data was evaluated in Study 1. This study compared design-based, weighted design-based, model-based, and weighted model-based approaches in a noninformative sampling design. First, results showed that the model-based approaches produced unbiased indirect effect estimates and smaller standard errors. Second, ignoring sampling weights led to substantial bias in the design-based approaches. Finally, in the model-based approaches, weighted parameter estimates and standard errors differed moderately from unweighted results. The model-based approaches were thereby suggested for testing multilevel mediation effects in large-scale, complex survey data. In addition, researchers were always encouraged to apply sampling weights in analysis. The advantages of applying sampling weights in model-based approach were less obvious when cluster sizes were large, and particularly when ICC was small.

The pursuit of evaluating various goodness-of-fit indices for testing measurement invariance has been a focus over the past decade. Study 2 expanded the investigation in MSEM. ICC and between-group difference accounted for a large proportion of variance in the model fit change. Among five model fit indices investigated in this study (i.e., χ^2,

CFI, RMSEA, SRMR, and TLI), ΔCFI and ΔSRMR in the level-specific approach had identical results to that of the standard approach. $\Delta SRMR_B$ appeared to be the most sensitive to noninvariant factor loadings among all criteria. $\Delta SRMR_B$ performed equally well in examining lack of intercept invariance when between-group difference was large. ΔRMSEA was less sensitive. Fractional changes in ΔCFI and ΔTLI indicated that neither was sensitive regardless of the level-specific approach or the standard approach. $\Delta\chi^2$ was able to detect noninvariant intercepts when between-group difference was large, whereas only detected noninvariant factor loadings when both ICC and between-group difference were large. In conclusion, level-specific $\Delta SRMR_B$ was suggested as a major index for examining between-level factor loading and intercept invariance in MSEM. $\Delta\chi^2$ can be a supplementary index.

DEDICATION

To my family,

and those who helped me throughout my research work

ACKNOWLEDGEMENTS

I would like to thank my committee co-chair, Dr. Oi-man Kwok and Dr. Myeongsun Yoon, and my committee members, Dr. Victor W. Willson and Dr. Wen Luo for their guidance and support throughout the course of this research.

Special thanks to Dr. Oi-man Kwok for his advice and guidance during my graduate study at Texas A&M University, which leads me to explore this dissertation topic. As a Chinese poem written by Fu Du (a Chinese poet in Tang Dynasty) says,

The good rain knows its season,

When spring arrives, then it comes.

It follows the wind secretly into the night,

And moistens all things softly, without sound. ("Welcome Rain at a Spring Night" 1-4. It was first translated by Hung, W., (1952) in *Tu Fu: China's Greatest Poet*.). Like a good rain at a spring night, Dr. Kwok always gives me helpful advice and generous encouragement. I am indebted to Dr. Kwok for his constant assistance, guidance and the tremendous support and opportunities he provided throughout my doctoral studies at Texas A&M University.

I would like to thank Dr. Jorge Gonzalez, Dr. Victor W. Willson, Dr. Quentin Dixon, and Dr. Lisako McKyer, Dr. Shanna Hagan-Burke for leading me in the path of doing research in several research projects. The opportunities to work with them, their wonderful expertise, valuable advice, and generous support are invaluable to me. Thanks also go to my friends and colleagues and the department faculty and staff for making my

time at Texas A&M University a great experience. They include, but are not limited to: Qiong Zhou, Russell Warner, Wenhua Lu, Eunsook Kim, MyungHee Im, Eunju June, Shuang Wu, Laura Frame, Heather Davis, Amber Wilmeth, Candice Hollub, Yunhee Bae, Mark Lai, Ning Ling, Nicola Ritter, Tammy Ramos, Cathy Watson, Kristie Stramaski, and Dr. Bob Hall. I also want to extend my gratitude to Dr. Eun Sook Kim at University of South Florida, who generously shared her simulation codes with me.

Finally, thanks to my family for their encouragement, patience and love. Life could not be a good journey without love.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

TABLE Page

TABLE Page

CHAPTER I

INTRODUCTION

Increasingly over the past two decades, applications of multilevel structural equation modeling (MSEM) have spread throughout the behavioral, educational, and social science research. Significant issues in techniques and applications of MSEM yet need our continuous attention, not only in substantive research but also in methodological area. This dissertation revisited two current issues in MSEM. One was the application of sampling weights in large-scale, complex data. The other was the goodness-of-fit indices for testing factorial invariance in multilevel data. Specifically, Study 1 aimed to assess the effects of using sampling weights on testing multilevel mediation effects in large-scale, complex survey data. Study 2 evaluated the performance of level-specific goodness-of-fit indices to test factorial invariance in MSEM. The two studies were two self-contained studies. Brief introductions about the two studies were presented below, respectively.

Study 1 addressed issues arising from special sampling designs used to collect large-scale, complex survey data (i.e., complex survey sampling). As is well known, conventional statistical methods assume the simple random sampling (SRS). Two properties of SRS are: 1) sampling units are independent from each other, and 2) each unit in the population has equal probabilities to be selected into the sample. Nevertheless, large-scale, complex survey data are typically collected via complex sampling designs, e.g., cluster sampling, multistage sampling, stratified sampling,

oversampling, and/or unequal probabilities of selection (Cochran, 1963; Longford, 1995; Snijders, 2012). The use of conventional statistical methods (e.g., t-test, ANOVA, and OLS regression) to analyze large-scale, complex survey data is seriously problematic due to the SRS violation. Therefore, taking into account sampling-related designs is important when analyzing large-scale, complex survey data. Substantive research has shown that ignoring the complex sample design would result in biased estimates and misleading inferences (e.g., Asparouhov, 2004; Pfeffermann, et al., 1998; Snijders, 2012; Stapleton, 2002).

In large-scale, complex survey data, sampling weights are available to researchers who do secondary data analysis. Objectively, sampling weights are intended to adjust individual unit's unequal probability of selection, to compensate for non-response rate, and/or to serve for post-stratification adjustment. With adjustments based on sampling weights, the selected sample is representative of the population of interest. Several simulation studies (e.g., Asparouhov, 2004, 2005, 2006; Cai, 2013; Kaplan & Ferguson, 1999; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002, 2006) demonstrated that incorporating of sampling weights in estimation can produce unbiased parameter estimates and standard error estimates. In practice, however, sampling weights were not always included when large scale data were analyzed (e.g., Bodovski, et al., 2013; Georges, et al., 2012; Gravers & Wright, 2011; Matthews & Kizzie, 2010; Peugh, 2013; Wright, et al., 2014). Two possibilities explain researchers' hesitation in using sampling weight when analyzing large-scale, complex survey data. First, the effects of sampling weights are not straightforward to researchers. In addition, there are various weighting

methods that can be used in analyzing large-scale, complex survey data and choosing the right weights is not an easy task. Second, a model-based approach, researchers believe, manages issues due to complex survey designs. However, this is not the case. The application of sampling weights remains an issue of significant concern to researchers using large-scale, complex survey data.

The abovementioned issues in analyzing large-scale, complex survey data need to be further investigated in new contexts. For example, though the importance of sampling weights has been highlighted in single-level analysis in literature, the role of sampling weights in MSEM is yet to be systematically explored. Further complicating matters, the pattern of sampling weights can be mixed in multilevel models. For example, the individual-level sampling weights are informative whereas the cluster-level sampling weights are noninformative, or vice versa. There are still arguments on how to include sampling weights in the estimation of multilevel models. To date, the effects of incorporating sampling weights on parameter estimates, standard errors estimates, and impact on power to detect multilevel mediation effects have not been addressed. In summary, the objective of Study 1 is to examine the effects of application of sampling weights on testing multilevel mediation effects in large-scale, complex survey data.

Study 2 is concerned with testing factorial invariance in multilevel data. In literature, testing measurement invariance has been intensively discussed in the context of single-level SEM regardless of in single-group CFA, multiple group CFA, multiple indicator multiple cause analysis (MIMIC), or restricted factor analysis (RFA) (Oort, 1992, 1998). Nevertheless, testing measurement invariance in MSEM becomes more

complicated than in single-level SEM. Some studies have pointed out that invariance testing involves parameters at multilevel data levels (i.e., the within- and between-level models separately) (Jak, Oort, & Dolan, 2013; Kim, Kwok, & Yoon, 2012; Ryu, 2014b). Measurement invariance evaluation thereby needs further investigation given new contexts, substantively and methodologically.

Over the past decade, evaluating the performance of various goodness of fit indices has been a focus for measurement invariance testing (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Through examining 20 goodness-of-fit indices, Cheung and Rensvold (2002) reported that ΔCFI, ΔGamma Hat, and ΔMcDonald's Noncentrality Index (Mc NCI) outperformed other ΔGFIs. Chen (2007) expanded the investigation to noninvariant settings (e.g., noninvariance in factor loading, intercepts, and residual variances). It was concluded that ΔSRMR was more sensitive to detect noninvariance in factor loadings than in intercepts and residual variances, while ΔCFI and ΔRMSEA equally performed under similar conditions. Consistent with Cheung and Rensvold's study, Meade, Johnson, and Braddy (2008) suggested to report ΔCFI and ΔMcDonald's Noncentrality Index for measurement invariance testing.

To date, all goodness-of-fit indices discussed previously in measurement invariance literature have been examined in single-level SEM framework. The performance of those model-fit indices in MSEM needs reinvestigations. Ryu and West (2009) examined the sensitivity of goodness-of-fit indices to detect model misspecification in MSEM. They reported that two commonly used fit indices (i.e., RMSEA and CFI) failed to detect between-level model misspecification. Consistent with

Yuan and Bentler's (2007) study, Ryu and West found that standard model fit indices (overall model fit) failed to capture misfit at the higher level model. The authors explained that because the cluster sizes were often larger than the number of clusters, the overall model fit was dominated by the within-level model. In a recent study by Hsu, Kwok, Lin and Acosta (2015), they found that CFI, TLI, and RMSEA were only sensitive to detect within-level model misspecification. In summary, standard goodness-of-fit indices (i.e., general model fit indices) reflected the fit of the overall model, whereas misfits at the specific level model, especially higher level model, were hard to be detected.

Study 2 is concerned with the sensitivity of model fit statistic to the model misspecification due to measurement noninvariance in MSEM. As discussed above, standard goodness-of-fit indices (for the entire model) may fail to detect noninvariance in specific level models. Given that noninvariance can occur in the within-level model only, in the between-level model only, or in both levels, the performance of various goodness-of-fit indices needs to be further investigated. In summary, the objective of Study 2 is to examine the level-specific goodness-of-fit indices in testing factorial invariance in MSEM. Meanwhile, the level-specific model fit evaluation is compared to the standard model fit approach.

CHPATER II

LITERATURE REVIEW

Social science, psychological, and educational researchers often encounter analysis of data of hierarchical structure. A common example in educational setting is given as follows. Students (level 1) are usually organized, or nested, in classrooms (level 2). Consider that students in the same classroom have the same teacher, curriculum, and complete similar tasks and activities. In this scenario, students in the same classroom tend to be more like one another than others from different classrooms on many attributes (e.g., learning outcomes). Other examples included employee nested in the organizations, voters from the same neighborhoods, patients visiting the same clinics, and English language learners who take the ESL classes. In all examples above, issue of observation dependency is obvious. In other words, the variability is smaller due to the correlations among observations. When analyzing data collected from hierarchical clusters, the correlation among observations needs to be taken into account. Given conventional statistical methods usually assume observation independency (i.e., independent sample), they are not suitable for analyzing data of clustered structure. Previous studies showed that ignoring data dependency often resulted in biased estimates of standard errors (overestimation) and misleading inferences (Hox, 2010; Rabe-Hesketh & Skrondal, 2006; Raudenbush & Bryk, 2002; Skinner, 1989).

Multilevel structural equation modeling (MSEM) has been widely used for analyzing data of hierarchical structure since a decade. As a general analytic framework,

MSEM integrates both multilevel modeling (MLM) techniques and SEM techniques (Kline, 2011; Mehta & Neale, 2005; Muthén & Asparouhov, 2012). On one hand, with the MLM tradition, MSEM can model relationships within clusters, but also between clusters from partitioning the total observed variance into within-level and between-level components. On the other hand, with the SEM tradition, MSEM can model means and covariance structures among multivariate data (Mehta & Neale, 2005). In addition, as a latent variable technique, measurement errors can be estimated. Muthén (1994) discussed MSEM for hierarchical data, such as multilevel factor and path models. However, only in the past two decades have analytical and computational advancements made MSEM popular among more interdisciplinary studies. Nowadays, well developed SEM packages, e.g., LISREL (Jöreskog & Sörbom, 1996), EQS (Benlter, 2004), AMOS (Arbuckle, 2006), and M*plus* (Muthén & Muthén, 1998–2012) allow researchers to apply MSEM with flexibilities.

A good understanding of measurement models is fundamental to understand MSEM. In this dissertation, a brief review of single-level measurement models was introduced, followed by a discussion extended to multilevel measurement models. In single-level measurement models, for individual $i$the relationship between a common factor (η_i) and the observed score (y_i) is examined. The general equation is

$$y_i = \upsilon + \Lambda\eta_i + \varepsilon_i, \quad (1)$$

where υ is the intercept vector;Λ is the factor loading matrix; and ε_i is the residual vector. As shown in the equation, a latent variable η_i is specified, which represents the

latent construct. One of the important assumptions in measurement models is that η_i and ε_i are uncorrelated.

In multilevel setting, suppose that individuals (i = 1, …, N) are drawn from J clusters (j = 1, …, J). To illustrate a multilevel measurement model, the subscript j (i.e., cluster) is introduced into Equation (1). The multilevel factor models are expressed in two equations (Rabe-Hesketh, Skrondal, & Pickles, 2004)

$$y_{ij} = \mu_j + \Lambda\eta_{ij} + \varepsilon_{ij}, \tag{2}$$

$$\mu_j = \mu + \Lambda\eta_j + \varepsilon_j, \tag{3}$$

where the latent factor means are treated as random effects varying across J clusters. The common factors (η_{ij}) and the unique factors (ε_{ij}) are assumed uncorrelated. In the multilevel measurement models, the latent factor means (η_{ij}) are divided into a within-level (i.e., Level 1) component and a between-level (i.e., Level 2) component, as shown by (Heck, 2001; Muthén, 1994)

$$\eta_{ij} = \alpha + \eta_{Bj} + \eta_{Wij}, \tag{4}$$

where α is the grand mean of η_{ij}. Latent random factors η_{Bj} and η_{Wij} represent between-cluster and within-cluster variation, respectively (Muthén, 1994). The expected values of η_{Bj} and η_{Wij} are zero. Similarly, the total factor variance can be decomposed into the between-cluster (ψ_B) and within-cluster (ψ_W) factor variance:

$$V(\eta_{ij}) = \psi_T = \psi_B + \psi_W, \tag{5}$$

where ψ_T denotes the total factor variance. Likewise, the residual variance (ε_{ij}) of the latent factor is decomposed into two components (i.e., Θ_B and Θ_W represent the between- and within- level residual variance, respectively) as

$$V(\varepsilon_{ij}) = \Theta_B + \Theta_W. \quad (6)$$

Similarly, the covariance structure for y_{ij} in Equation (2) is then given as

$$V(y_{ij}) = \Sigma_T = \Sigma_B + \Sigma_W, \quad (7)$$

where Σ_W and Σ_B refer to the within (level-1) and between (level-2) covariance matrix, respectively. Till this point, the covariance and mean structure for two-level factor models can be summarized as (Heck, 2001; Jak, Oort, & Dolan, 2014):

$$\Sigma_B = \Lambda_B \psi_B \Lambda_B{}' + \Theta_B, \quad (8)$$

$$\mu_B = \tau_B + \Lambda_B \alpha_B, \quad (9)$$

$$\Sigma_W = \Lambda_W \psi_W \Lambda_W{}' + \Theta_W. \quad (10)$$

Equation (8) and (10) describe the covariance structure between groups and within groups, respectively. In Equation 9, α_B denotes common factor means at the between level. It should be noted that no mean structure at the within level is specified. It is because that individual scores are defined as cluster means plus individual deviations from the cluster means at the between-level. Therefore, the expected values of within-level intercepts are zero and the overall mean equalsμ_B.

In chapter II, the concepts of multilevel measurement models were presented. Two MSEM models were considered in this dissertation. One was a multilevel mediation model and the other was a multilevel measurement model. A more detailed review of each model was discussed in the following chapters.

CHAPTER III

STUDY 1: THE EFFECTS OF SAMPLING WEIGHTS ON TESTING MULTILEVEL MEDIATION EFFECTS IN LARGE-SCALE, COMPLEX SURVEY DATA

LARGE-SCALE, COMPLEX SURVEY DATA

Large-scale, complex survey data are usually large amount of data. National survey data and international assessment data are two major types. National survey data are collected by government agencies such as the National Center for Education Statistics (NCES). International large-scale assessments data (e.g., the Programme for International Student Assessment (PISA) are organized by international organizations, for example, the Organization for Economic Co-operation and Development (OECD). Regardless of national survey data or international assessment data, as it is known, they are large-scale, complex survey data based on complex sample designs. The complex sample designs include stratification, clustering, multi-stage sampling, and unequal probability of selection of observations, and/or nonresponse and post-stratification adjustment.

Complex sample designs, as described above, are special features of large-scale, complex survey data. On one hand, those techniques have been shown to increase efficiency in survey sampling, to save money and time. On the other hand, those survey techniques used make survey analyses complicate. Conventional statistical methods assume SRS, in which each observation unit is randomly selected from the population and selected with equal probability of selection. The use of complex survey sampling

results in unequal probabilities of selection and data non-independency issue. Under those circumstances, conventional statistical methods are no longer appropriate. Taking into account complex sample designs is necessary in the survey analysis of large-scale, complex survey data.

To address every effect of data characteristics on model estimation due to complex sample designs is difficult. In addition, the effect could be mixed. Such effects include stratification, clustering, unequal sampling, and finite population (Lumley, 2004). In this study, two major issues of complex sample design were addressed, including: issues of data dependency and unequal probabilities of selection. In the next two sections, sampling weights were first presented which adjust unequal probabilities of selection. Second, design-based and model-based approaches were introduced which account for data dependency.

SAMPLING WEIGHTS

Sampling weights are available to researchers who do secondary data analysis of large-scale, complex survey data (e.g., national survey data and international assessment data). The incorporation of sampling weights in modeling data based on complex sampling designs has been highly recommended by researchers (e.g., Asparouhov, 2004, 2005, 2006; Cai, 2013; Kaplan & Ferguson, 1999; Korn & Graubard, 1995; Lee et al., 1989; Rabe-Hesketh & Skrondal, 2006; Skinner, 1989; Stapleton, 2002, 2006, 2012). Despite all these recommendations of using sampling weights, substantive researchers didn't include sampling weights in analysis as a regular practice.

Sampling weights contain important information of unequal probability of selection adjustment, non-response adjustment, and/or post-stratification adjustment due to complex sampling designs. The incorporation of sampling weights in analysis functions as necessary adjustments in survey data estimation. In simple random sampling scheme, every element in the population has the same probability of being selected in the sample. However, based on complex sampling designs (e.g., proportional sampling), sample units do not have equal probability of selection included in the sample. Therefore, conventional statistics methods are not appropriate for analyzing complex survey data. Some simulation studies found that ignoring unequal probability of selection resulted in biased parameter estimate in single-level SEM and multiple regression analysis (Kaplan & Ferguson, 1999; Korn & Graubard, 1995; Lee et al., 1989), additionally, biased standard errors estimates in multilevel analysis (Asparouhov, 2004; Cai, 2013).

Furthermore, the type of sampling design should be a concern for researchers, that is, whether the sampling design is informative or noninformative. According to Snijders (2012), informative design is defined that "the distribution of the residuals is affected by the sampling design" (p. 222). If the ways that units are selected with unequal probabilities of selection are related to variables that researchers are interested (e.g., oversampling a specific group and the grouping variable is the outcome variable), the sampling design is said to be informative. As evaluated in some simulations studies, it was important to include sampling weights in the estimation when sampling design is informative (e.g., Asparouhov, 2006; Cai, 2013; Grilli, et. al., 2004; Pfeffermann, 1993, 1998; Snijders, 2012; Stapleton, 2002). Researcher reiterated the necessity of

incorporating this kind of sampling designs in modeling and it was pointed out that not only the parameter estimates but also standard errors of parameter estimates may be seriously biased in the informative sampling design if weights were not applied (e.g., Binder &Roberst, 2001; Cai, 2013; Jia et el., 2012; Preffermann, et al., 1998; Sugden & Smith, 1984). Muthén and Satorra (1995) proposed two approaches for alleviating the impact of informative designs in model estimations. One was to include the design weights. The other way was to include the design variables when specifying the model, thus the correlations of the design variables to the other variables were controlled in the hypothesized model. While researchers have limited information of the design variables in the large-scale, complex survey data, the application of sampling weights is a good alternative. When the sampling design is noninformative, which is defined that "the distribution of residuals is independent of the sampling design" (Snijders, 2012, p. 222), Korn and Graubard (1995) suggested to incorporate effective sampling weights or higher-level weights in the estimation. However, Snijders (2012) argued that the use of sampling weights was redundant given the sampling design was noninformative. Because there are still arguments on the application of sampling weights and also on the weighting method, further research is needed to investigate the impact of sampling weights under various situations. In the next section, two kinds of sampling weights were introduced.

Types of Sampling Weights

Raw Weights. When data are collected with unequal probabilities of selection (e.g., proportional sampling), selection bias occurs. In this case, sample data need to be

weighted to adjust for selection bias. One frequently used sampling weights is raw weight which the summation adds up to the population size (i.e., N). Raw weights are commonly used in descriptive analyses (Kalton, 1989), which are defined as:

$$\sum_{i=1}^{n} \omega_i = N, \quad (1)$$

where $w_i = \frac{1}{p_i}$ and p_i is each individual's inclusion probability. Note that N is the population size rather than the sample size. Though parameter estimates are less affected by N, standard errors and fit statistics, particularly the maximum likelihood-ratio chi-square statistic, are very sensitive to the N being used (Kaplan & Ferguson, 1999). Researchers had proposed other types of weighting methods. Potthoff, Woodbury and Manton (1992) described other two types of weights, that is, normalized weights (also called relative weights in some studies) and effective weights, which were discussed in the following sections.

Normalized Sampling Weights. Normalized sampling weights (also called relative weights, cf. Longford, 1995; Pothoff et al., 1992; Stapleton, 2002; Thomas & Heck, 2001) sum to the actual sample size n (rather than the population N), defined as:

$$\sum_{i=1}^{n} \omega_i^* = n. \quad (2)$$

In the simplest case, if p_i = n/N and $\omega_i^* = \frac{n}{N} \times \frac{1}{p_i} = \frac{n}{N} \times \frac{N}{n} = 1$, the sample weights sum to the sample size n ($1 \times n = n$). Based on Equation 1-2, $\frac{1}{p_i}$ is the raw weight w_i, as defined previously in Equation (1). $\frac{n}{N}$ can be viewed as the normalizing factor. Potthoff, Woodbury, and Manton (1992) described the procedures of normalization. That is,

rescale the raw weights to the normalized weights by scaling the raw weights by a constant λ, defined as:

$$\lambda = \frac{n}{\sum_{i=1}^{n} w_i} = \frac{n}{N}, \quad (3)$$

where w_i is the raw weight for individual i. In other words, the normalized sampling weights can be obtained by multiplying the raw weights with the normalizing factor.

In a two-stage cluster sampling, i denotes the individual level unit and j notates the cluster level unit. $\omega_{ij}^{*} = \frac{n_j}{n} \times \frac{N_i}{N} \times \frac{1}{p_j p_{i|j}}$ wherep_j is the probability of a cluster being selected and $p_{i|j}$ is the probability of an individual i included in that cluster j. $\frac{n_j}{n}$ is the ratio of the number of selected clusters n_j to the total number of clusters n in the population. $\frac{N_i}{N}$ is the ratio of the number of observations N_i selected in cluster j to the total sample N_j of cluster j. $\frac{n_j}{n} \times \frac{N_i}{N}$ can be viewed as the normalizing factor. After the normalization process, ω_i^{*} sums to the total sample size. A simple example below is given illustrating raw sample weight and normalized sampling weight based on one-stage stratified sampling design. A balanced sample of ten is equally drawn with five from two subpopulations, such as 2,000 and 8,000 in each pool. The two subpopulations constitute the population of 10,000. The normalized sampling weights from the pool of 2,000 and 8,000 are 0.4, and 1.6, respectively, whereas the raw sample weights are 400 and 1600, respectively. Obviously, the normalized sampling weights sum up to 10 (sample size) but the raw sample weights sum to 10,000 (the population size).

Because standard error estimates are sensitive to the n used, different types of sampling weights may result in huge difference. Some studies (e.g., Asparouhov, 2006;

Grilli&Pratesi, 2004; Kaplan & Ferguson, 1999; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2012) proved that applying raw weights tended to underestimate standard errors and inflate Type I errors. In statistical software, for example, SPSS, M*plus* and LISREL, the normalized weighting method is implemented as the default. For example, it is set up by using the CLUSTER setting in M*plus* and the CLUSTER command In LISREL. Effective weights have also been studied by some researchers (e.g., Pfeffermann, Skinner, et al., 1998; Stapleton, 2002) but they are out of scope of this dissertation and for details please read Stapleton's (2002) study.

Sampling Weights in Multilevel Analysis

As discussed previously, when sampling weights are informative, sampling weights should be included in analysis. In addition, normalizing scaling method is often recommended. However, the application of sampling weight in multilevel analysis is more complicated than in single-level analysis. In multilevel analysis, Asparouhov (2004) and Snijders (2012) suggested that when sampling weights at both the individual level and cluster level were noninformative, sampling weights were not needed in the analysis. In contrast, when both the individual-level and cluster-level sampling weights were informative, the application of sampling weights was necessary (Aaparouhov, 2004). Pfeffermann, Skinner, et al. (1998) compared three methods of weights (i.e., raw weights, normalized weights, and effective weights) and concluded that the application of sampling weights resulted in unbiased estimates regardless of which method used. They recommended to use the normalized weights for which reflected the actual sample

size. Nevertheless, Korn and Graubard (1995) argued that effective weighting method (effective sampling weights) should be used at either level.

In addition to the arguments on weighting methods, further complicating matters was there were mixed patterns of sampling weights. That is, the cluster-level sampling weights were noninformative but the individual-level sampling weights were informative, and vice versa. In the case that only individual-level weights were informative, Pfeffermann, Skinner, et al. (1998) proposed that scaling the individual-level sampling weights for each cluster using the effective weighting method (effective weights). They claimed that it produced unbiased estimates for both variance components. Arguably, Asparouhov (2004) suggested to scale the weights within each cluster using the normalized weighting method. When informative sampling weights were present at cluster level only, Asparouhov (2004) pointed out that a single-level weighted analysis produced unbiased results. The author explained that only the cluster level units with unequal probabilities of selection won't affect the outcome variable at the individual level. Though there was no agreement on the choice of weighting method, one consensus was when the sampling weights at the individual level were informative, sampling weights should be scaled and included in the multilevel analysis.

ANALYTICAL APPROACHES FOR ANALYZING COMPLEX SURVEY DATA

Issues of data dependency should be a concern in analyzing data collected from complex sampling designs (e.g., cluster sampling or multi-stage sampling). Design-based and model-based approaches are two popular alternatives for analyzing complex

survey data (Cai, 2013; Kalton, 1983; Snijders, 2012; Wu & Kwok, 2012). Meanwhile, incorporating sampling weights into multilevel analysis was also highly recommended for analyzing complex survey data (Asparouhov, 2006; Graubard & Korn, 1996; Grilli & Pratesi, 2004; Kovacevic & Rai, 2003; Pfeffermann et al., 1998). Both design- and model-based approach take the multilevel data structure of large-scale, complex survey data into account. Differently, the design-based approach specifies a single-level model while making appropriate standard error and fit statistics adjustment (e.g., a sandwich estimator) to account for data dependency. The model-based approach requires level-specific models for each data level (i.e., multilevel models). From modeling random effects (e.g., random intercepts and/or slopes) that vary across clusters, in other words, partitioning the population variance into within- and between-level variance, the model-based approach produces unbiased parameter estimates and standard error estimate. In summary, design-based approaches produce unbiased parameter estimates and correctly-adjusted standard errors (Hahs-Vaughn, 2005; Stapleton, 2002). Nevertheless, the model-based approach has been shown as a promising approach given its flexibility for specifying multilevel models at different data levels, partitioning the total variance into within and between parts, and resulting in unbiased parameter and standard error estimates (Hox, 2010; Hox & Kreft, 1994; Kaplan & Elliott, 1997; Muthén, 1994; Muthén & Satorra, 1995; Stapleton, 2002). In the next section, the two approaches were discussed and the application of sampling weights in estimation was also presented.

Design-based Approaches

The design-based analysis methods include a single-level model specification and appropriate standard errors and chi-square test statistics adjustments. In M*plus*, MLR (maximum likelihood with robust standard errors and chi-square test statistics) estimation method is employed by using the TYPE = COMPLEX routine (default estimation method when observed dependent variables are all continuous). Standard errors are computed using a sandwich estimator, which are robust to non-normality of outcomes and observation independency.

When samples are selected with unequal probabilities of selection, sampling weights contain important information of selection probabilities. The pseudo maximum likelihood (PML) method is used for estimation in the single level model with sampling weights incorporated. PML is now widely implemented in M*plus*, SAS, and STATA and other commercial software packages for addressing unequal probabilities of selection. In the design-based approaches, researchers can specify the weights variable to allow for weights incorporation in estimation. The PML parameters of interest are obtained by maximizing the weighted likelihood function as given by:

$$l_n(\theta) = \sum_{i=1}^{n} \omega_i \log f_\theta(x_i), \quad (4)$$

where ω_i is the sampling weights and subscript *i* notates individual observations 1, … , *n*.

The robust sandwich estimator is used for standard error adjustment under model misspecification assumption (e.g., observation independence here). When x is IID random variable $x_1, \ldots x_n$, with parameter θ, the asymptotic variance is

$$\hat{J}_n(\hat{\theta}_n)^{-1}\hat{V}_n(\hat{\theta}_n)\hat{J}_n(\hat{\theta}_n)^{-1}. \quad (5)$$

The weights can also be plugged in the "sandwich estimator". For example, in a linear regression, when weights incorporated the variance is given by

$$V(\hat{B}) = (X'WX)^{-1}V[\sum_{i \in s} w_i x_i'(y_i - x_i'\hat{B})](X'WX)^{-1}, \quad (6)$$

where w_i represents sampling weights and the subscript *i* runs from 1 to *n* for all observations in the sample. W is the weight matrix. The use of sandwich standard error estimator allows for necessary adjustments of standard error estimates, which is robust to data nonnormality and observation nonindependence.

Model-based Approaches

Compared to the design-based approach in which only a single level model is specified, the model-based approach allows researchers to model both within- and between-relations with models specified at each level. Particularly, researchers can study the effects of higher level variables on the outcome variables at individual level. In M*plus*, using the TYPE = TWOLEVEL routine, a multilevel model with random effects (e.g., random intercepts and/or slopes) is estimated (Muthén & Muthén, 1998–2012). Similar to the design-based approach, MLR (maximum likelihood with robust standard errors and chi-square test statistics) and "sandwich estimator" are used for parameter and standard error estimation. Both design-based and model-based approaches produce unbiased standard error estimates. Differently, the model-based approach partitions the variance of the observed dependent variable into within-component and between-component, while design-based approach makes adjustments to standard error estimates based on the sampling designs.

When sampling weights are applied in model estimation, the multilevel pseudo maximum likelihood (MPML) method is used. This estimation method can be used to estimate any general multilevel model (Asparouhov, 2004). Similar to maximum likelihood method, the MPML parameter estimates are obtained by maximizing the weighted likelihood function. Define θ_1 and θ_2 as the parameters for the individual and cluster level; w_{ij} and w_j as the level 1 and level 2 weights. The weighted likelihood function is

$$l(\theta_1, \theta_2) = \prod_j (\int (\prod_i f(y_{ij}|x_{ij}, \eta_j, \theta_1)^{w_{ij}s_{1j}}) \phi(\eta_j|x_j, \theta_2) d\eta_j)^{w_j s_{2j}}, \quad (7)$$

where cluster j runs 1, … to M and individual iis 1, …, n_j; s_{1j} and s_{2j} are level 1 and level 2 weights scaling factors (Asparouhov, 2004). Standard errors are computed using the *sandwich estimator*. With sampling weights, the asymptotic covariance of the MPML estimates takes the form of

$$(L'')^{-1} \left(\sum_j (s_{2j} w_j)^2 L_j' L_j'^T \right) (L'')^{-1}, \quad (8)$$

in which $l_j = \int (\prod_i f(y_{ij}, x_{ij}, \eta_j, \theta_1)^{w_{ij}s_{ij}}) \phi(\eta_j, x_j, \theta_2) d\eta_j$. L' and L'' refer to the first and second derivative of the log-likelihoods. w_j is the weight of the top level unit and s_{2j} is the level 2 weights scaling constant (Asparouhov, 2004).

Till this point, it was presented that design-based and model-based approach have employed different estimation methods in the application of sampling weights in estimation. The MPML estimation method required that information of sampling weights should be available at the individual level and cluster level.

MULTILEVEL MEDIATION MODELS

In a single-level simple mediation analysis, the mediation model is expressed as the following regression equations:

$$M_i = i_M + aX_i + e_{M_i}, \quad (9)$$

$$Y_i = i_Y + c'X_i + bM_i + e_{Y_i}, \quad (10)$$

where Y is the dependent variable, X is the independent variable, and M is the mediator (MacKinnon, 2008). The terms i_M and i_Y denote intercepts; e_{M_i} and e_{Y_i} represent residuals. Coefficient a is the effect of X on M, coefficient c' is the direct effect of X on Y, and coefficient b describes the effect of M on Y while controlling for X. ab is termed as the indirect effect (i.e., X on Y through M). The total effect of X on Y equals to the indirect effect (i.e., ab) plus the direct effect (i.e., c').

In educational and organization settings, researchers often encounter data of multilevel structure. For example, the dependent variable (Y) and the independent variable (X) are level-1 variables, but the mediator (M) is a level-2 variable. An empirical example is that X is the amount of homework a student does (level-1), Y is a student's score on the exam (level-1), and M is a type of educational program that a classroom is assigned (level-2) to where the student attends. The mediation effect occur across levels. In this case, simple mediation model is no more appropriate but multilevel mediation analysis should be used. Bauer, Preacher, and Gill (2006) summarized previous studies (e.g., Kenny, Kashy, & Bolger, 1998; Krull & MacKinnon, 1999, 2001) and illustrated three types of multilevel mediation models, which were discussed as follows.

In multilevel mediation models, i denotes individual and j represents clusters. In a "2 – 1 – 1" multilevel mediation model, X is a level-2 predictor. M and Y are level-1 variable. The effect of M on Y can be a random effect, varying between level-2 units. The model equations are given by:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}M_{ij} + e_{ij}, \quad (11)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}X_j + U_{0j}, \quad (12)$$

$$\beta_{1j} = \gamma_{10} + U_{1j}. \quad (13)$$

It should be noted that the effect of M on Y is estimated as random (U_{1j}). However, this model can be specified as a fix effect of M on Y (exclude the term of U_{1j}).

In a "1 – 1 – 1" multilevel mediation model, all three variables (X, M, and Y) are measured at level 1. All effects are allowed to vary across level-1 units. That is, the effect of both X on M (not shown in the equation) and M on Y can be random effects. The indirect effect of X on Y through M is also estimated as random. The model equations are given as follows:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \beta_{2j}M_{ij} + e_{ij}, \quad (14)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + U_{0j}, \quad (15)$$

$$\beta_{1j} = \gamma_{10} + U_{1j}, \quad (16)$$

$$\beta_{2j} = \gamma_{20} + U_{2j}. \quad (17)$$

Similarly, this model can have a fix effect of M on Y (exclude the term of U_{2j}), a fix effect of X on Y (exclude the term of U_{1j}), or a fix effect of X on M (not shown in the equation). When either one of two effects is estimated as fixed, no covariance term between the path coefficients of the two effects is needed. When both are estimated as

random, a covariance term needs to be included, for example, $\hat{\sigma}_{a_j b_j}$ represents the covariance between the two random effects (i.e., the effect of X on M and the effect of M on Y).

In a "2 – 2 – 1" multilevel mediation model, X and M are measured at level 2 and Y is a level-1 outcome. All effects (X on M, M on Y, and X on Y) are estimated as fix effects. The model equations include:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + e_{ij}, \tag{18}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}X_j + \gamma_{02}M_j + U_{0j}. \tag{19}$$

In all three types of multilevel mediation models discussed above, the effect of X on M is not shown in the equations. In this study, a "2 – 1 – 1" multilevel mediation model was used with no random effect was specified, which was discussed in a later section.

Testing Multilevel Mediation Effects

In Preacher and Hayes (2008), the authors provided a thorough review on the theory of testing mediation effects. The formal definition of mediation was put forward by Baron and Kenny (1986) in a simple mediation model, which was expressed as the difference between the total effect and the direct effect of X on Y. Preacher and Hayes summarized five statistical approaches to assess mediation, including: Baron and Kenny's (1986) casual steps strategy, partial correlation strategies, differences in coefficients strategies, nested model strategy, and product of coefficients strategies. Among them, three approaches (i.e., casual steps strategy, differences in coefficients strategies, and product of coefficients strategies) are used very often in mediation analysis. Preacher and Hayes (2008) described their pros and cons and suggested not to

use the partial correlation strategies and the nested model strategy. MacKinnon, Fairchild, and Fritz (2007) proposed that using the product of coefficients (i.e., $\hat{a} \times \hat{b}$) or difference in coefficients (i.e., $\hat{c} - \hat{c}'$) yielded similar results for testing mediation effects. However, they also warned that the two methods were only exchangeable in some circumstances (e.g., single-mediator model) but not everywhere (e.g., multilevel models, logistic or probit regression, and survival analysis). In this study, the product of coefficients method was adopted. The standard error for the parameter estimated (i.e., $\hat{a}*\hat{b}$) is defined as (MacKinnon, 2008, p. 92):

$$s_{\hat{a}\hat{b}} = \sqrt{\hat{a}^2 s_{\hat{b}}^2 + \hat{b}^2 s_{\hat{a}}^2}, \tag{20}$$

where $s_{\hat{a}}$ and $s_{\hat{b}}$ are estimated standard errors for $\hat{a}$ and $\hat{b}$, respectively.

The Sobel test (Sobel, 1982) is used to test for significance of the indirect effects, defined as:

$$t = \frac{\hat{a}\hat{b}}{s_{\hat{a}\hat{b}}}, \tag{21}$$

in which a normal distribution of t statistics is assumed. MacKinnon (2008, p. 94) suggested using the distribution of the product (i.e., $\hat{a}*\hat{b}$) to test for significance if normality assumption is violated. Either in design-based or model-based approaches, testing mediation effects is specified in the same way in M*plus* (e.g., by specifying a new parameter, that is, $\hat{a}*\hat{b}$).

In the discussion above, mediation effects were addressed in single-level models. In the following sections, testing mediation was discussed in multilevel mediation models. As Kenny et al. (2003), Bauer, Preacher, and Gil (2006), and Bullock et al.

(2010) demonstrated, if $\hat{a}$ and $\hat{b}$ slope were both random, that is, the two slopes varied across clusters, correlations between the two random coefficients also occurred at the higher level. The indirect effect was then expressed as:

$$\hat{a}\hat{b} + \hat{\sigma}_{a_j,b_j}, \tag{22}$$

where $\hat{\sigma}_{a_j b_j}$ represented the covariance between the two random effects in lower level multilevel mediation models (i.e., all variables were measured at low level – Level 1). Note that both $\hat{a}$ and $\hat{b}$ are random slope coefficients, varying across the higher level units. Similarly, with a covariance term the variance of the indirect effect is given as (Kenny et al., 2003):

$$\hat{a}^2\sigma_{b_j}^2 + \hat{b}^2\sigma_{a_j}^2 + \sigma_{a_j}^2\sigma_{b_j}^2 + 2\hat{a}\hat{b}\sigma_{a_j,b_j} + \sigma_{a_j,b_j}^2. \tag{23}$$

The Equations (22) and (23) are only applicable to 1→1→1 lower level multilevel mediation models in which random effects are assumed for both coefficients (see Bauer et al., 2006, pp. 143-144; Preacher & Selig, 2010). In upper level mediation (i.e., 2→2→1) and lower level mediation (i.e., 2→1→1) models, no covariance term between slopes is needed because no concurrent random effects are assumed (Preacher & Selig, 2010). In other words, only one random slope is modeled at a time (e.g., either $\hat{a}$ or $\hat{b}$, but not $\hat{a}$ and $\hat{b}$ simultaneously).

Up till now, I have summarized the theoretical framework of Study 1 including two kinds of weighting method, two analytical approaches to analyze complex survey data, and testing multilevel mediation models in the multilevel data. In the next, the

incorporation of sampling weights and how it affected estimates in multilevel mediation models was examined through a Monte Carlo simulation.

Though in literature researchers have discussed the importance of incorporating sampling weights in single-level and multilevel analysis, the effects of using sampling weights in multilevel mediation models were still unknown. In other words, the impact of including sampling weights on estimating indirect effect parameters, standard errors, and testing mediation effects was to be examined. Taken all together, the purpose of this study was to examine the effect of sampling weights on estimating multilevel mediation models using four methods (i.e., unweighted design-based, design-based weighted, unweighted model-based, and model-based weighted). The results obtained from this study were twofold for researchers: 1) choosing between design- and model-based approaches, and 2) how to treat sampling weights in testing multilevel mediation effects in large-scale, complex survey data.

METHOD

It was of interest to examine the impact of utilizing sampling weights on testing multilevel mediation effects in large-scale, complex survey data in MSEM framework. In this study, a simple two-level mediation model was chosen with a Level-2 predicator influencing a Level-1 mediator which in turn affected a Level-1 outcome. For example, a study by Schulting et al. (2005) using the ECLS-K data examined the effect of school-based kindergarten transition policies and practices (a Level-2 predicator) on child

academic outcomes (a Level-1 outcome), as mediated by parent-initiated involvement in school (a Level-1 mediator).

The equations for the superpopulation model used for data generation were shown as follows:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}M_{ij} + e_{ij}, \quad (24)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}X_j + U_{0j}, \quad (25)$$

$$\beta_{1j} = \gamma_{10}, \quad (26)$$

where subscripts *i* and *j* referred to individuals and Level-2 units (e.g., clusters), respectively; γ_{00} was the ground mean (mean across all clusters); γ_{10} was the fix effect for the mediator variable M_{ij}; e_{ij}, and U_{0j} were the Level-1 and Level-2 residuals (i.e., random effects), respectively. As also depicted in Figure 3.1, X was a level-2 predictor and the mediator M and the outcome variable Y were at level-1. In the two-level mediation model, the effect of X on Y (a between indirect effect) was of research interest. It was notable that in equations 24-26 the effect of mediator on outcome can be a random effect (i.e., a random slope term U_{1j} in equation 26), but was treated as fixed effect in this study for simplicity. The results from this model without random slopes were applicable to other types of mediation models (see Preacher, Zhang, & Zyphur, 2011).

Data Generation and Analyses

The sampling scheme designed in this study was a one-stage stratified sampling, with disproportionate sampling of observations in different strata reflecting oversampling. Empirical samples with oversampling are common in large-scale,

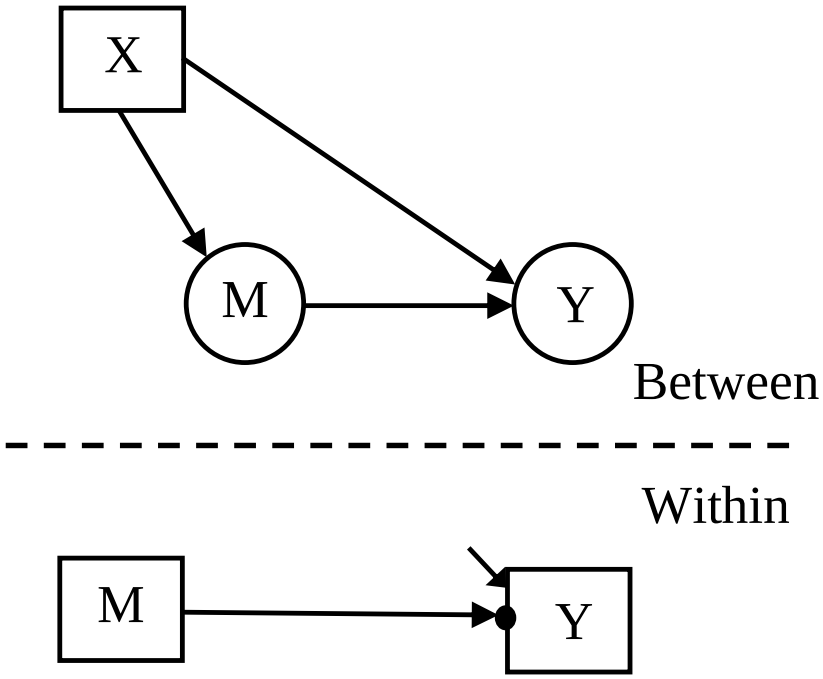

FIGURE 1 Two-level Mediation Model.

complex survey data, for example, the oversampled black and Hispanic persons in the 1995-2004 National Health Interview Survey (National Center for Health Statistics, 2000), the oversampled full-time females, Black non-Hispanics and Hispanics, Asian/Pacific Islanders in the National Study of Postsecondary Faculty (Selfa et al., 1997), and the oversampled of Blacks, Hispanics, military, disadvantaged Whites in the National Longitudinal Survey of Youth, 1979 Cohort (Bureau of Labor Statistics, 2012), to name a few.

To compute sampling weights, samples were manipulated to draw from several strata (not all) of different sizes according to: 1) the population included several strata. When the same number of observations drawn from strata of different sizes, unequal probabilities of selection incurred, 2) the design assumed that the population size and

strata sizes were known. Therefore, sampling weights can be calculated based on units' inclusion probabilities.

Three experimental factors that impacted on testing mediation effects in complex survey data were considered. Specifically, the impact of cluster sample sizes, the size of the intraclass correlation coefficient (ICC), and the effect size of the indirect effect were examined. In all, the simulation used a 2 (two levels of mediation effect) × 2 (two levels of ICCs) × 3 (three levels of cluster sizes) factorial design with a total of 12,000 replications. The design factors were chosen based on studies on multilevel modeling, the application of sampling weights, and mediation effects (e.g., Hox & Maas, 2001; Kaplan & Ferguson, 1999; Kenny, Korchmaros, & Bolger, 2003; Krull & MacKinnon, 2001; Stapleton, 2002). The details of these experimental conditions were as follows:

Cluster Size: 10, 20, and 50,

Effect Size of Mediation: low mediation (.09) and high mediation (.36),

Level of ICC: ICC low or high (i.e., .1 or .5).

The superpopulation constituted 100,000 observations and was expected to be a combination/mixture of two substrata of different sizes. For example, one stratum had 30,000 observations while the other stratum included 70,000 cases.

First, strata data were generated separately and every stratum constituted several clusters. The large stratum was expected to have more clusters/groups than the small stratum. Clustered data were structured in a two-level hierarchical manner (e.g., two-level data with individuals nested within clusters). The cluster size was set to be a constant (i.e., 1000 observations per cluster). Note that the cluster size was pretty large

in this study because the next step was a disproportional sampling from the clusters. The substrata models were assumed to hold in each stratum as given in equations 23–25. The parameters in the substrata model were set as follows: 1) Path a (from X to M) was set equal to path b (from M to Y). They equaled to .6 and .3 for high and low mediation, respectively, 2) The population variance of X was fixed at 1; variances of M and Y were manipulated to obtain different ICC, that is, $ICC_M = ICC_Y = .1$ or .5, 3) Intercepts were set to 0, and 4) The indirect effect from X to Y is a*b (i.e., $a*b = 0.3 \times 0.3 = 0.09$ and $a*b = 0.6 \times 0.6 = 0.36$ for low and high mediation, respectively). The substrata data were generated using M*plus* (Muthén & Muthén, 1998-2012) and then combined to form the superpopulation.

Second, cluster samples were drawn from each of the two strata within their respective clusters. Note that a SRS of individual units in multilevel data destroyed the hierarchically structured data for multilevel modeling (Roberts & Fan, 2004; Wang, Xie, & Fisher, 2011), so it was manipulated that the disproportionate sampling of observations from different strata were drawn within the clusters. The number of observations from the small stratum was oversampled compared to the large stratum. The sampling procedures and sampling weights computations were conducted in SAS (SAS 9.3). Normalized sampling weights were calculated and attached to each observation in all data sets. Take an example, if 600 observations were drawn from stratum 1 which had 30,000 cases the normalized weight was.65. If 700 observations were drawn from stratum 2 which included 70,000 cases the normalized weight was 1.3.

Without applying sampling weights in estimation, stratum 1 was certainly overrepresented in the sample.

Last, each simulated data set was then forwarded to M*plus* to be analyzed using different methods, including: (a) unweighted design-based, (b) design-based weighted, (c) unweighted model-based, and (d) model-based weighted. Given that sampling weights can be incorporated in the estimation, the effects of sampling weights on testing mediation effects were evaluated from comparing weighted with unweighted analyses. Three outcomes including parameter estimates, sampling variability, and goodness-of-fit statistics were summarized to explore the impact of sampling weights in testing multilevel mediation effects.

Note that the superpopulation was a mixture of two substrata, and the parameter values used to generate the substrata data were changed when the two strata merged together. Therefore, in the result section I won't compare the results obtained from the simulation to the initial parameter values used for data generation. Rather, the results were compared to unbiased estimates of the superpopulation parameters which was a combination of the two strata.

The results were summarized in terms of three criteria: 1) the accuracy for which the indirect effect was estimated, specifically, percent underestimation or overestimation was calculated representing the relative bias in parameter estimates (i.e., 100×(estimate average - population value)/population value), 2) the efficiency about the estimated indirect effect around the population indirect effect, particularly, the root mean squared error (RMSE) of the estimated indirect effect was computed, expressed as a function of

empirical standard deviation (ESD) and bias (i.e., RMSE = $\sqrt{ESD^2 + BIAS^2}$, Preacher, Zhang, & Zyphur, 2011), and 3) the adequacy (model adequacy) with respect to the effects of sampling weights on model fit indices including AIC, BIC, SABIC, CFI, RMSEA, and SRMR.

RESULTS

Accuracy: The Indirect Effect Parameter Estimates

Results of relative percentage bias were reported in Table 3.1. It can be seen that the model-based approaches greatly outperformed the design-based approaches, indicated by the fact that in all conditions (i.e., high mediation high ICC, high mediation low ICC, low mediation high ICC, and low mediation low ICC), the model-based approaches resulted in substantially smaller bias than that for the design-based approaches. This trend remained with the change of cluster sizes. The second finding was that ignoring sampling weights led to larger bias, as can be seen by comparing the weighted and unweighted results (design-based unweighted versus design-based weighted and model-based unweighted versus model-based weighted). With increasing cluster sizes, the pattern was consistent except for the conditions of cluster size at 20. When the indirect effect was small and ICC was high, the unweighted design-based approaches had the most and unacceptable bias. The model-based weighted approaches had smaller biases. To sum up, the model-based approach resulted in less bias with respect to the design-based approach. In addition, the application of sampling weights improved the indirect effect estimates.

TABLE 3.1 Percentage Relative Bias of the Indirect Effect Using Different Methods

n	DB	DB+W	MB	MB+W
High Mediation High ICC				
10	6.08	2.66	.31	.21
20	6.08	3.62	.03	.39
50	6.03	3.42	-.05	.00
High Mediation Low ICC				
10	1.21	.98	.17	.03
20	1.34	1.50	.22	.36
50	1.11	1.11	-.06	-.03
Low Mediation High ICC				
10	11.80	4.75	.57	.47
20	11.80	5.08	.57	.75
50	11.80	4.92	.19	.19
Low Mediation Low ICC				
10	3.07	2.53	.33	.22
20	3.20	3.20	.33	.56
50	2.93	2.80	.00	.11

Note. n: cluster size.
DB: design-based unweighted.
DB+W: design-based weighted.
MB: model-based unweighted.
MB+W: model-based weighted.
ICC = Intra-class correlation.

Efficiency: Empirical Standard Deviation and RMSE of the Estimated Indirect Effect

According to a criterion reported in Preacher, Zhang, and Zyphur (2011), larger values of empirical standard deviation (ESD) indicated lower efficiency. As shown in Table 3.2, the model-based approaches exhibited higher efficiency than the design-based approaches with smaller ESDs, only except in high mediation low ICC and low mediation low ICC conditions. The results suggested that the model-based approaches produced smaller standard errors than the design-based approaches. Small changes were observed in ESDs when cluster sizes increased. The second finding was that the weighted conditions showed advantages over the unweighted analysis for the design-based approaches. In contrast, the application of sampling weights in the model-based approaches was less efficient. In other words, no much difference was observed between unweighted and weighted model-based analyses.

Based on the RMSE results (see Table 3.3), the model-based approaches noticeably showed to be more efficient than the design-based approaches under all conditions with the smallest RMSE values. When cluster sizes increased, RMSEs decreased to some extent. The incorporation of sampling weights resulted in smaller RMSE values. In low mediation high ICC condition, unweighted design-based approaches tended to be the least efficient with the largest RMSE values. Decreased efficiency was noted for low mediation low ICC conditions at larger cluster sizes (e.g., over 20) in the model-based weighted analysis. In summary, in terms of efficiency the model-based approaches showed to be promising with smaller standard errors and RMSEA.

Adequacy

To simplify presentation, only results of high mediation low ICC condition and low mediation low ICC condition were presented (see Table 3.4 and Table 3.5) in terms of model adequacy (i.e., information criteria and goodness-of-fit indices). Though results of high ICC conditions were not summarized here, similar findings were obtained while comparing the design-based approach with the model-based approach, and comparing the weighted with the unweighted analysis.

Given smaller information criteria (e.g., AIC, BIC, and SABIC) indicated better model fits, the model-based approaches appeared superior to the design-based approaches in high mediation low ICC condition (see Table 3.4). In addition, according to goodness-of-fit indices (e.g., RMSEA, CFI, TLI, and SRMR) the model-based approaches resulted in better mode fit. Furthermore, the application of sampling weights improved the model fits regardless of in the design-based approaches or the model-based approaches. Meanwhile, as the cluster sample sizes increased, some fit indices (e.g., AIC, BIC, SABIC, and chi-squares) increased.

Similar trends were identified in low mediation low ICC condition (see Table 3.5). According to fit indices including AIC, BIC, and SABIC, the model-based approaches were still better than the design-based approaches and the application of sampling weights resulted in better fits in both approaches. When cluster sizes increased, AIC, BIC, and SABIC turned larger while small changes in RMSEA, CFI, TLI, and SRMR were observed.

Summary of Results

The study had two objectives including the comparison between the design- and model-based approaches and the evaluation of the application of sampling weights in a noninformative design. The main findings from this simulation study were as follows in terms of accuracy, efficiency, and model adequacy: (a) The unweighted design-based approaches was the most biased method, resulting in relatively larger bias of the indirect effect estimates, less efficiency, and worse fit compared to the other three alternatives in any condition (i.e., weighted design-based, model-based, and weighted model-based). The application of sampling weights was one way that reduced estimation bias in the design-based approach; (b) The model-based approaches were preferred in testing multilevel mediation effects in large-scale, complex survey data relative to the design-based approaches, because, as expected, the model-based approaches resulted in less bias, higher efficiency, and better model fit; (c) Differences between weighted and unweighted analyses revealed that the application of sampling weights brought advantages, even though it was less explicit in the model-based approach when cluster sizes were large, and particularly when ICC was small. Based on the above findings, recommendations and suggestions for practice were thereby given in the discussion section.

TABLE 3.2 Empirical Standard Deviation of the Estimate of the Indirect Effect Using Different Methods

n	DB	DB+W	MB	MB+W
High Mediation High ICC				
10	.053	.048	.030	.031
20	.051	.046	.026	.025
50	.050	.045	.020	.020
High Mediation Low ICC				
10	.023	.024	.024	.025
20	.017	.017	.018	.018
50	.012	.012	.012	.012
Low Mediation High ICC				
10	.025	.023	.016	.017
20	.024	.022	.014	.014
50	.024	.022	.013	.013
Low Mediation Low ICC				
10	.008	.008	.008	.008
20	.006	.006	.006	.006
50	.005	.005	.004	.004

Note. n: cluster size.
DB: design-based unweighted.
DB+W: design-based weighted.
MB: model-based unweighted.
MB+W: model-based weighted.
ICC = Intra-class correlation.

TABLE 3.3 Root Mean Square Error of the Estimate of the Indirect Effect Using Different Methods

n	DB	DB+W	MB	MB+W
High Mediation High ICC				
10	6.080	2.660	.311	.212
20	6.080	3.620	.040	.391
50	6.030	3.420	.054	.020
High Mediation Low ICC				
10	1.210	.980	.172	.039
20	1.340	1.500	.221	.360
50	1.110	1.110	.061	.032
Low Mediation High ICC				
10	11.800	4.750	.570	.470
20	11.800	5.080	.570	.750
50	11.800	4.920	.190	.190
Low Mediation Low ICC				
10	3.070	2.530	.330	.220
20	3.200	3.200	.330	.560
50	2.930	2.800	.004	.110

Note. n: cluster size.
DB: design-based unweighted.
DB+W: design-based weighted.
MB: model-based unweighted.
MB+W: model-based weighted.
ICC = Intra-class correlation.

TABLE 3.4. Maximum Likelihood Goodness-of-Fit Statistics for *High Mediation Low ICC*

Fit Statistics	Cluster Sizes			Cluster Sizes		
	10	20	50	10	20	50
		Design-based			*Weighted design-based*	
AIC	9062.31	18114.95	45278.24	9060.41	18113.81	45277.00
BIC	9093.33	18150.13	45318.92	9091.43	18148.99	45317.68
SABIC	9074.27	18131.06	45299.85	9072.37	18129.92	45298.61
CHISQ	89.04	109.30	125.92	101.72	138.35	167.11
RMSEA	.255	.201	.138	.273	.225	.159
CFI	.827	.828	.831	.815	.810	.808
TLI	.482	.484	.492	.445	.423	.423
SRMR	.080	.080	.080	.081	.081	.081
		Model-based			Weighted model-based	
AIC	8694.14	17273.73	42923.12	8688.08	17269.85	42918.40
BIC	8735.50	17320.63	42977.36	8729.44	17316.76	42972.64
SABIC	8710.09	17295.21	42951.94	8704.03	17291.34	42947.22
CHISQ	3.65	3.26	3.54	3.08	2.44	2.61
RMSEA	.015	.009	.006	.013	.008	.004
CFI	.999	.999	1.000	.999	1.000	1.000
TLI	.998	.999	1.000	.999	1.000	1.000
SRMR (within)	.001	.000	.000	.001	.000	.000
SRMR (between)	.047	.031	.019	.050	.033	.020

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC. CHISQ = chi-square; RMSEA = root mean squared error of approximation; CFI = comparative fit index; TLI = the Tucker-Lewis index; SRMR = standardized root mean square residual.

TABLE 3.5 Maximum Likelihood Goodness-of-Fit Statistics for *Low Mediation Low ICC*

Fit Statistics	Cluster Sizes			Cluster Sizes		
	10	20	50	10	20	50
		Design-based			*Weighted design-based*	
AIC	5602.83	11197.09	27984.38	5600.70	11195.60	27982.17
BIC	5633.86	11232.27	28025.06	5631.72	11230.78	28022.85
SABIC	5614.80	11213.21	28005.99	5612.66	11211.72	28003.78
CHISQ	14.42	16.89	18.99	16.60	20.54	24.01
RMSEA	.099	.077	.052	.107	.085	.059
CFI	.966	.978	.987	.958	.973	.985
TLI	.898	.933	.961	.875	.919	.956
SRMR	.045	.045	.045	.046	.045	.046
		Model-based			Weighted model-based	
AIC	5480.44	10886.34	27030.23	5474.43	10882.46	27025.48
BIC	5521.80	10933.24	27084.47	5515.79	10929.36	27079.72
SABIC	5496.39	10907.82	27059.05	5490.37	10903.95	27054.30
CHISQ	4.23	3.16	6.58	4.033	2.53	2.40
RMSEA	.017	.010	.006	.013	.008	.004
CFI	.995	.998	.998	.995	.998	.999
TLI	.992	.998	.995	.993	.999	1.000
SRMR (within)	.001	.000	.000	.002	.001	.000
SRMR (between)	.046	.031	.019	.047	.032	.020

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC. CHISQ = chi-square; RMSEA = root mean squared error of approximation; CFI = comparative fit index; TLI = the Tucker-Lewis index; SRMR = standardized root mean square residual.

DISCUSSION

Sampling weights have increasingly been applied when analyzing large-scale, complex data. However, there were still doubts in its use among applied researchers due to several reasons. First, multiple sampling weights were available in large-scale, complex survey data and choosing the appropriate sampling weights was complicated (e.g., weights at different levels in the data). Second, there were also several weighting methods (e.g., raw sampling weights, normalized sampling weights, and effective sampling weights). Several simulation studies addressed the scaling issue (c.f., Stapleton, 2002, 2006 for discussion on raw weights, relative weights, and effective weights). However, it was unclear which weighting method was appropriate given certain circumstances. Third, after researchers decided to use specific sampling weights, choosing the most appropriate analytical approach posed difficulties because different approaches may cause divergent results (e.g., the design-based versus model-based approaches).

Findings from this study suggested that the model-based approaches were preferred to the design-based approaches and the application of sampling weight application was also recommended for testing mediation effects in large-scale, complex survey data. If researchers chose the design-based approaches, the application of sampling weights was necessary. The findings were consistent to previous simulation studies (see, e.g., Asparouhov, 2004; Pfeffermann, 1996). For larger cluster sizes (e.g., over 20), particularly when ICC was small (e.g., 0.1), the advantages of applying sampling weights were less obvious in the model-based approaches. These findings

provided useful suggestions to applied researchers who targeted large-scale, complex survey data analysis.

As with most simulation studies, results were limited to the study design which needed further investigations. First, this study employed only a simple proportional sampling design and sampling weights reflected the probabilities of selection in different strata. In practice, unequal probabilities of selection can be very complicated (e.g., multistage sampling). In addition, sampling weights in large-scale, complex survey data can be used for multiple adjustments, for example, nonresponse and post-stratification adjustments. Therefore, further research was needed addressing the complexity of sampling schemes and various adjustments.

A second limitation was that the superpopulation model used in this study was a lower level multilevel mediation model to study the impact of sampling weights in testing multilevel mediation effects. In practice, researchers may specify different types of multilevel mediation models. Therefore, caution should be taken generalizing potential effects of sampling weights on testing multilevel mediation effects.

Nevertheless with these limitations, this study highlighted two important searches: 1) modeling large-scale, complex survey data under MSEM framework, 2) how sampling weights affected testing multilevel mediation effects in a noninformative sampling design. In literature, studied more addressed informative sampling designs. For example, Pfeffermann et al. (1998) focused on multilevel models in informative designs and Asparouhov (2005) also considered latent variable modeling in informative sampling designs. In conclusion, this study suggested to use a model-based approach

when testing multilevel mediation effects in large-scale, complex survey data, and the application of sampling weights was also encouraged.

CHAPTER IV

STUDY 2: EVALUATION OF THE LEVEL-SPECIFIC GOODNESS-OF-FIT INDICES IN TESTING FACTORIAL INVARIANCE IN MULTILEVEL DATA

MEASUREMENT INVARIANCE

This literature review of testing factorial invariance in MSEM included: 1) the importance of testing measurement invariance, 2) statistical methods of testing measurement invariance (e.g., multiple-group CFA, MIMIC, or Bayesian SEM), 3) model fit evaluation in SEM, and 4) testing factorial invariance in multiple group multilevel factor analysis.

Why Does Measurement Invariance Matter

Measurement invariance (MI), as known as measurement equivalence, is theoretically defined that the conditional probability to attain an observed score is identical for individuals with given ability regardless of group membership (Mellenbergh, 1989; Millsap, 2011; Yoon & Millsap, 2007). In other words, when MI assumed, a measure assesses the same attribute in the same way across groups. Vandenberg & Lance (2000) described the context that "the term groups refers either to independent groups or to the same group measured longitudinally on multiple occasions" (p. 10). In Millsap (2011), it was emphasized that MI implied the important psychometric property of a measure (i.e., unbiased measurement).

Measurement invariance (MI) was usually assumed rather than tested in some empirical studies. Researchers tended to simply assume MI before making group

comparisons. For example, a simple t-test assumes MI. In advanced analytic techniques, e.g., latent growth curve modeling, MI is also assumed. Nevertheless, measurement invariance can be violated due to many reasons and different settings. Specific examples include cross-cultural studies, in which group comparisons are made across countries or ethnic groups (e.g., Chen, 2008; Dolan et al., 2004; van de Vijver, 2009), longitudinal studies, in which multiple groups are compared across time points for possible changes (e.g., Pentz & Chou, 1994), focal studies on language acquisition, in which bilingual or multilingual students are compared to monolingual ones using the same measure (e.g., De Houwer et al., 2014). In all these cases, testing measurement invariance should be emphasized before comparing across groups.

A significant amount of literature has demonstrated the importance of testing measurement invariance (e.g., Borsboom, 2006; Byrne & Watkins, 2003; Chen, 2007; Chen, Sousa, & West, 2005; Cheung & Rensvold, 2002; Horn & McArdle, 1992; Meredith, 1993; Millsap, 2011; Reise, Widaman, & Pugh, 1993; Schumitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Some studies evaluated the impact of noninvariance (e.g., lack of invariance and partial invariance). Specific examples included, just to name a few: Millsap and Kwok (2004) provided evidence of the impact of partial invariance on selection accuracy via a Monte Carol study; Chen (2008) examined the impact of lacking factor loading and intercept invariance on slope and factor mean comparisons in three simulations; Through an empirical example, Schimitt et al. (2011) investigated the impact of measurement noninvariance on construct correlations, mean differences, and

relations. Testing measurement invariance has been highlighted in methodological research. Nevertheless, testing measurement invariance should be further advocated to applied researchers.

When testing measurement invariance (MI), a stepwise approach of examining four levels of MI, proposed by Widaman and Reise (1997) who synthesized prior research (e.g., Horn, McArdle, & Mason, 1983; Jöreskog, 1971; Jöreskog & Sörbom, 1999; Meredith, 1964a, 1964b, 1993), has been well established in social science research. Widaman and Reise discussed four levels of measurement invariance, including: configural invariance, weak invariance, strong invariance, and strict invariance. Strictly, only when invariance at all levels is validated, measurement invariance is said to be fully fulfilled. However, though it is desirable, full invariance is not easy to be established in practice. Partial invariance is then proposed with mixed invariant and noninvariant parameters (some model parameters are invariant, but not all, see Byrne, Shavelson, & Muthén, 1989). Nevertheless, though partial invariance draws increasing attention (see Schumitt & Kuljanin's (2008) review), the controversies and debates are still going on. One reason is that no established protocols in psychometrics on how to manage partial invariance have been proposed. In this dissertation, the approach of partial invariance was not discussed.

Approaches of Testing Measurement Invariance

Structural equation modeling (SEM), widely applied in the social science research, is frequently used for measurement invariance testing. Under a SEM framework, the multi-group confirmatory factor analysis (multi-group CFA) and

multiple-indicators multiple-causes (MIMIC, Jöreskog & Goldberger, 1975) modeling are two popular alternatives. Arguably, CFA method is the most popular method in SEM literature. Virtually all studies reviewed by Vandenberg and Lance (2000) focused on CFA method. In another review by Schimitt and Kuljanin (2008), 85.23% of empirical studies (75 out of 88 reviewed articles) reported using CFA to examine measurement invariance. A literature search in PsycINFO (EBSCO) database involving quantitative and empirical studies published since 2000 was conducted. At least 1,000 peer reviewed articles reported using the CFA method, whereas 45 studies applied the MIMIC method, which evidenced the generality of the CFA method in substantive area. One reason of its popularity is that multi-group analysis is straightforward (i.e., separate models are specified for each group simultaneously).

Though sometimes CFA and MIMIC methods are used alternatively they have some differences. Kim, Yoon and Lee (2012) pointed out that MIMIC assumes equal factor loadings and intercepts, so it was vulnerable to test weak invariance. In different factor models, CFA and MIMIC perform unequally in terms of power and accuracy. For example, in the first-order factor models, Hancock, Lawrence, and Nevitt (2000) reported very little difference between the two approaches under a balanced design. Multi-group CFA was yet concluded to be more preferable as sample sizes between groups became more disparate. In second-order factor models CFA models were recommended with less Type I error in a balance design (Zou, 2009). MIMIC method also has advantages. When the grouping variable is continuous, the MIMIC method is more efficient without having to creating many subgroups. When the research interests

are to test all levels of measurement invariance, multiple-group CFA is recommended with a step-by-step procedure.

An alternative approach, Bayesian structural equation modeling (Bayesian SEM) to test measurement invariance draws attention with advancements in Bayesian analysis, e.g., the increasing computational availability of Markov chain Monte Carlo (MCMC) techniques. With unique strengths and accessible complex computation, Bayesian SEM has been applied in psychology and management research to test measurement invariance (e.g., Fong & Ho, 2013; Golay et al., 2012; Schoot, et al., 2013; Zyphur & Oswald, 2013).

In nature, Bayesian SEM is still a confirmatory factor analysis in the SEM framework, whereas the implement of Bayesian estimation provides advantages over standard fitting of SEMs. Most popular software packages, e.g., LISREL (Jöreskog & Sörbom, 1999), EQS (Benlter, 2004), AMOS (Arbuckle, 2006), and M*plus* (Muthén & Muthén, 1998-2012) provide estimation methods such as maximum likelihood (ML) and generalized least squares (GLS). The use of ML estimation in multi-group CFA or MIMIC modeling represents the approach of frequentist practice. As it is known, ML estimation is the most popular method in the current SEM literature. However, with advancements in computing power, the growth of Bayesian analysis has been a phenomena. Muthén and Asparouhov (2012) illustrated the flexibility of Bayesian approach Bayesian SEM compared to ML-based CFA or MIMIC modeling. Next a brief introduction of Bayesian SEM for testing measurement invariance was presented.

Using multiple group CFA to test measurement invariance, equal parameter

constraints are imposed (e.g., zero loadings for non-invariant items in single-group CFA and equal parameter constraints across groups either in factor loading, intercept, or in residual variance). Muthén and Asparouhov (2012) introduced the concept of "approximate measurement invariance". In single-group CFA the non-invariant items are specified as zero mean with small variance. In multi-group CFA, equal parameter constraints across groups are also relaxed in the Bayesian approach (i.e., zero mean difference but with small variance). As pointed out by some studies (e.g., Cole, et al., 2007; Kolenikov, 2011), too many parameter restrictions could easily lead to poor model fit and substantial bias in parameter estimates. Therefore, with relatively relaxed model constraints (e.g., approximate measurement invariance) the Bayesian approach shows to be less restrictive than the conventional CFA (e.g., equal constraints).

The disadvantages of the approximate measurement invariance are also brought into arguments. Jak (2014) illustrated some limitations of Bayesian SEM in a recent application. One issue was with respect to the specification of prior distribution because different priors resulted in divergent results. Muthén and Asparouhov (2012) recommended approximate invariance with zero-mean, small-variance priors, but the small-variance specification yet needs further evidence under various conditions (e.g., prior variance ranges from 0.01 to 0.1 in their study). In addition, the heavy computational loads of MCMC methods are also problematic. Jak (2014) reported nonconvergence problems when the number of groups was large (e.g., 156 groups). In conclusion, the BSEM approach is a new tool and useful, but further research is needed.

MODEL FIT EVALUATION IN SEM

When assessing overall fit of a SEM model, it is common that researchers use the likelihood ratio test (LR test) and alternative goodness-of-fit indices (GFIs). As it is known, the chi-square difference statistics is sensitive to sample size. Therefore, a variety of goodness-of-fit indices has been applied as model fit measures. Among them, CFI, RMSEA, SRMR, and TLI are the most commonly used indices reported by researchers in social science. Guidelines of cut-off criteria for those fit measures were proposed in Hu and Bentler (1999), which have been widely used in SEM literature.

When testing measurement invariance, some researchers applied the LR test in which the difference between the model chi-square statistics (ΔX^2) was computed. Other researchers estimated a series of nested models (i.e., the constrained models and the unconstrained models) and compared the GFIs of those models. For a constrained model, the between-group constraints were held equal across groups while an unconstrained model had all constraints freely estimated (no between-group constraints). In addition to ΔX^2, difference in goodness-of-fit difference statistics (ΔGFI) has also been used for testing measurement invariance. Some studies investigated the performance of various ΔGFI for testing measurement invariance (e.g., Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Those authors also presented cut-off points (critical values) for rejection of the null hypothesis of measurement invariance, which were summarized in Table 4.1. Some inconsistency among the recommendations of cutoff points (ΔGFI) was observed in these studies. For example, Cheung and Rensvold (2002) listed the first (or 99th) percentile values as the

critical values for ΔCFI, ΔGamma-hat, and ΔMcDonald's Noncentrality Index, that is, .01, .001, and .02, respectively. Slightly differently, Chen (2007) proposed criteria including: ΔCFI of -.005, ΔRMSEA of .01, and ΔGamma-hat of -.005, whereas Meade et al. (2008) proposed a general cutoff value of .002 for the ΔCFI and .008 for ΔMcDonald's Noncentrality Index. Taken together, the discrepancy in the cutoff criteria in previous simulation yet needs further research.

Over the last decade, the use of multilevel SEM has been rising across many disciplines. It is a question that whether the guidelines of single-level SEM model fit indices are still applicable to multilevel SEM models. Yuan and Bentler (2003, 2007) explored level-specific model fit evaluation (segregating approach) in multilevel SEM in which fit indices were obtained for each level (i.e., individual level and group level, respectively). Similarly, Ryu (2014a) and Ryu and West (2009) proposed the approach of level-specific model fit evaluation (partially-saturated approach). A consistent finding in those studies was that the standard model fit indices were measures of model fit for the overall model but may not be good for models at specific levels. Moreover, the standard model fit indices performed poorly to detect model misspecification at the higher model level.

In Hox's (2010) multilevel analysis book, he proposed model fit evaluation for multilevel factor models. He introduced the notion of saturate models that "the saturated model estimates all covariance between all variables. It has no degrees of freedom, and always fits the data perfectly (p. 307)". Based on this notion, model fit for the within-level models can be obtained by specifying a saturated model for the between matrix.

TABLE 4.1 A Summary of Cut-off Values for Rejection of Measurement Invariance in Previous Studies

Type	Cheung & Rensvold (2002)	α = .05, Chen (2007)	α = .01, Chen (2007)	Meade, Johnson, and Braddy (2008)
ΔCFI	change of ≤ -.01	change of ≤ -.005	change of ≤ -.005	change of ≤ -.002
ΔRMSEA	n.a.	change of ≤ .010	change of ≤ .010	n.a.
ΔGamma- hat	change of ≤ -.001	change of ≤ -.005	change of ≤ -.005	change of ≤ -.005
ΔMc NCI	change of ≤ -.02	change of ≤ -.010	change of ≤ -.010	change of ≤ -.008

Note. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; Mc NCI = McDonald's noncentrality index.

Similarly, with a saturated model specified for the within matrix, the fit represents the model fit at the between level. Following Hox's idea, Ryu and West (2009) developed the level-specific model fit indices for the between- and within-level models, respectively, for multilevel SEMs. The proposed goodness-of-fit indices included level-specific chi-square test statistics, level-specific CFIs, and level-specific RMSEAs (i.e., X_B^2, X_W^2, CFI_B, CFI_W, $RMSEA_B$, and $RMSEA_W$). Subscripts B and W denoted between-level and within-level model, respectively. Notations from Ryu & West (2009) were used thereafter in this study. In the two-level analysis, M*plus* (Muthén & Muthén, 1998-2012) also provides SRMR at the between and within level separately (i.e., SRMR-between and SRMR-within), but not for other indices (e.g., CFI, RMSEA, and TLI).

According to Ryu and West's (2009) simulation study, the level-specific model fit evaluation can be used to examine specific models at different levels separately. It performed equally well for detecting model misspecification at lower and higher level. In the next section, the level-specific model evaluation in MSEM was introduced.

Level-Specific Model Evaluation in MSEM

As it is known, M*plus* (Muthén & Muthén, 1998-2012) provides SRMR for the within-level model and for between-level model separately in M*plus* output. Nevertheless, other common indices, such as CFI, TLI, and RMSEA, only have one general value for the entire model. In this dissertation, following Ryu and West's (2009) study I explored the level-specific model fit indices in MSEM to test factorial invariance. As discussed previously, within-level model fit can be obtained by specifying a saturated between-level model. Similarly, between-level model fit can be

obtained by specifying a saturated within-level model. In M*plus*, a saturated model can be specified by correlating all observed dependent variables with each other. In the following sections, level-specific fit indices were discussed. All the equations and notations in this section were taken from Ryu and West's (2009) study if not specified.

Level-Specific Chi-Square Statistics. Level-specific chi-square statistics were first introduced because some other indices (e.g., CFI and RMSEA) were computed based on chi-square statistics. Given a saturated within-level model, the between-level chi-square test statistics are computed from

$$X_B^2 = F_{ML}\left[\sum_B(\hat{\theta}), \sum_W(\hat{\theta}_S)\right] - F_{ML}\left[\sum_B(\hat{\theta}_S), \sum_W(\hat{\theta}_S)\right], \quad (1)$$

where Σ_B is the between covariance structure. S denotes a saturated model. θ is a vector of model parameters. The lack of invariance at the between level is thus reflected by the discrepancy between $\sum_B(\hat{\theta})$and $\sum_B(\hat{\theta}_S)$. The df_B is the difference between the number of parameters in the saturated between-level model and the hypothesized between-level model. Similarly, the within-level chi-square test statistics are computed from

$$X_W^2 = F_{ML}\left[\sum_W(\hat{\theta}), \sum_B(\hat{\theta}_S)\right] - F_{ML}\left[\sum_W(\hat{\theta}_S), \sum_B(\hat{\theta}_S)\right]. \quad (2)$$

The df_W is the difference between the number of parameters in the saturated within-level model and the hypothesized within-level model. The lack of invariance at the within level is indicated in the difference between $\sum_W(\hat{\theta})$ and $\sum_W(\hat{\theta}_S)$.

Level-Specific Goodness of Fit Indices. In the following section, level-specific goodness-of-fit indices for four indices were discussed, including: CFI, RMSEA, SRMR, and TLI.

1. CFI

The comparative fit index (CFI) (Bentler, 1990) reflects the goodness of fit of the hypothesized model compared to an independence model. In the independence model, all the variances are freely estimated but all the covariances are constrained to be zero. In M*plus* (Muthén & Muthén, 1998-2012), it is achieved by specifying all off-diagonal elements zero in the variance covariance matrix.

To calculate the level-specific CFIs (i.e., $\mathrm{CFI_B}$ and $\mathrm{CFI_W}$), the level-specific chi-square statistics should be computed first. H denotes the hypothesized model and I represents the independence model. $X^2_{H,B}$ can be obtained by specifying a hypothesized between-level model while specifying a saturated model at the within level. The $df_{H,B}$ is the difference between the number of parameters in the hypothesized between-level model and the saturated between-level model. Similarly, $X^2_{I,B}$ can be obtained by specifying an independence model at the between level and a saturated model at the within level. The $df_{I,B}$ is the difference between the number of parameters in the independence between-level model and the saturated between-level model. Substitute the chi-square statistics and df into Equation (3), the CFI_B for the between level model is given by

$$CFI_B = (1 - \frac{Max[(X^2_{H,B} - df_{H,B}),0]}{Max[(X^2_{I,B} - df_{I,B}),0]}). \quad (3)$$

Similarly, with a saturated model at the between level, an independence and a hypothesized model at the within level, respectively, the CFI_W is given by

$$CFI_W = (1 - \frac{Max[(X^2_{H,W} - df_{H,W}),0]}{Max[(X^2_{I,W} - df_{I,W}),0]}). \quad (4)$$

2. TLI

Like CFI, TLI is used to compare the fit of the fitted model against an independence model that assumes that variables are uncorrelated. In this dissertation, I also included TLI (Tucker-Lewis index, Tucker & Lewis, 1973). On one hand, TLI is one commonly used index. It can be easily extracted from M*plus* output. On the other hand, TLI compensates for model complexity (Brown, 2006). The complexity penalty in TLI is given by the chi-square to DF ratio, which is more sensitive to model parsimony than CFI (Kenny, 2008).

To obtain the level-specific TLIs, the chi-squares and dfs need to be calculated first. They can be obtained in the same way as in calculating level-specific CFIs shown in Equations (3) (4). The level-specific TLIs are given by

$$TLI_B = \frac{\frac{X^2_{I,B}}{df_{I,B}} - \frac{X^2_{H,B}}{df_{H,B}}}{\frac{X^2_{I,B}}{df_{I,B}} - 1}, \tag{5}$$

$$TLI_W = \frac{\frac{X^2_{I,W}}{df_{I,W}} - \frac{X^2_{H,W}}{df_{H,W}}}{\frac{X^2_{I,W}}{df_{I,W}} - 1}. \tag{6}$$

3. RMSEA

The root mean square error of approximation (RMSEA) (Steiger, 1990) is a measure of the lack of fit of the model to the population covariance matrix. Like TLI index, RMSEA is sensitive to model parsimony and its penalty for model complexity is the chi square to df ratio. RMSEA is currently one of the most popular index (the others include CFI and SRMR) in SEM literature.

The level-specific RMSEAs ($RMSEA_B$ and $RMSEA_W$) can be computed based on corresponding chi-square test statistics and *dfs* derived from Equations 3-4. J and N denote the total number of clusters and cluster size, respectively. The Equations for $RMSEA_B$ and $RMSEA_W$ are

$$RMSEA_B = \sqrt{Max[\left(\frac{X_B^2 - df_B}{df_B(J)}\right), 0]}, \quad (7)$$

$$RMSEA_W = \sqrt{Max[\left(\frac{X_W^2 - df_W}{df_W(N-J)}\right), 0]}. \quad (8)$$

When the level specific RMSEAs are negative, zero is reported instead (Hox, 2010).

4. SRMR

The standardized root mean square residual (SRMR) (Bentler, 1995; Muthén, 2004) is a standardized measure of the difference between the observed covariance matrix and the model-implied covariance matrix. M*plus* provides $SRMR_B$ and $SRMR_W$ for model at specific level, respectively. $SRMR_B$ is given by

$$SRMR_B = \sqrt{\frac{2\sum_i \sum_j r_{ijB}^2}{p(p+1)}}, \quad (9)$$

where p is the number of observed variables, p(p+1)/2 is the number of unique elements in the between-level residual covariance matrix with *i* rows and *j* columns. r_{ijB} is a residual in a between-level residual covariance matrix ($\hat{\Sigma}_{rijB}$) which is defined

$$\hat{\Sigma}_{rijB} = \sum_B(\hat{\theta}) - \sum_B(\hat{\theta}_S). \quad (10)$$

The between-level residual covariance matrix is the difference between the estimated between-level covariance matrix and the saturated between-level covariance matrix while the within-level model is saturated. Similarly, with the saturated model at the

between level, $SRMR_W$ is given by

$$SRMR_W = \sqrt{\frac{2\sum_i \sum_j r_{ijW}^2}{p(p+1)}}, \quad (11)$$

where r_{ijW} is a residual in a within-level residual covariance matrix ($\hat{\Sigma}_{rijW}$) which is defined

$$\hat{\Sigma}_{rijW} = \sum_W(\hat{\theta}) - \sum_W(\hat{\theta}_S). \quad (12)$$

The within-level residual covariance matrix is the difference between the estimated within-level covariance matrix and the saturated within-level covariance matrix while the between-level model is saturated.

TESTING FACTORIAL INVARIANCE IN MSEM

In the single-level multiple-group CFA approach, four levels of measurement invariance (i.e., configural invariance, metric/weak invariance, scalar/strong invariance, and strict invariance) can be examined (see Widaman & Reise, 1997). For example, weak invariance presumes that factor loadings are equal across g (g = 1,...,G) groups.

$$\Lambda_1 = \Lambda_2 = \cdots = \Lambda_G \quad (13)$$

where Λ stands for the factor loading vector. Different levels of measurement invariance across groups can be tested in a hierarchical order as described below (from the least to the most restricted). First, configural invariance describes whether the measurement model holds across groups. Second, on the basis of configural invariance, weak invariance examines whether equality in factor loadings is further established across groups. Third, given that weak invariance holds, equality in intercepts across groups is

then constrained for testing strong invariance. Fourth, with established strong invariance, equality in unique variances across groups is further examined at the most restricted level, that is, strict invariance. In the next section, testing factorial invariance was reviewed in multilevel SEM models.

Within structural equation modeling, measurement invariance is known as factorial invariance framed in factor analytic terms (Byrne, Shavelson, & Muthén, 1989; Widaman et al., 2010; Widaman & Reise, 1997). Hereafter, factorial invariance was to be used in this study regarding invariance test in multilevel factor models. In the context of multilevel SEM, factorial invariance testing should be conducted both in the between-level model and the within-level model given that noninvariance can occur in the within level model and/or in the between level model. In the following sections, testing factorial invariance was discussed in two scenarios, that is, when the grouping variable is at the between level and when the grouping variable is at the within level.

Between-level Grouping Variable

When the grouping variable is at the between level, multiple group multilevel confirmatory factor analysis (multiple group multilevel CFA) can be used for testing factorial invariance in multilevel data. The multilevel models are formulated as follows (Kim, Kwok, &Yoon, 2012; Muthén, 1994)

$$y_{ijg} = \upsilon_{Bg} + \Lambda_{Bg}\eta_{Bjg} + \varepsilon_{Bjg} + \Lambda_{Wg}\eta_{Wjg} + \varepsilon_{Wjg}, \quad (14)$$

$$V(y_{ijg}) = \Sigma_{Bg} + \Sigma_{Wg}, \quad (15)$$

$$\Sigma_{Bg} = \Lambda_{Bg}\psi_{Bg}{\Lambda_{Bg}}' + \Theta_{Bg}, \text{ and} \quad (16)$$

$$\Sigma_{Wg} = \Lambda_{Wg}\psi_{Wg}{\Lambda_{Wg}}' + \Theta_{Wg}. \quad (17)$$

In Equations (14) to (17), y_{ijg} is the observed score of individual i (i = 1, …, N) nested within cluster j (j = 1, …, J) in group g (g = 1, …, G). The model parameters include intercept (υ), factor loading (Λ), latent factor (η), and the unique factor (ε).

Conceptually, factorial invariance testing in multiple group multilevel CFA is comparable to that in multiple group single-level CFA. In other words, the Widaman and Reise's (1997) approach can be readily adapted. Specifically, the identified four levels of factorial invariance in a hierarchical order are: 1) configural invariance, in which the pattern of factor loadings in both within- and between-level models is identical across groups, in other words, all groups can be fitted with the same multilevel factor models simultaneously, 2) weak factorial invariance, in which factor loadings of both within- and between-level models are invariant across groups (i.e., $\Lambda_{Bg} = \Lambda_B$ and $\Lambda_{Wg} = \Lambda_W$), 3) strong factorial invariance, in which intercepts of between-level models are equal across groups, that is, $\upsilon_{Bg} = \upsilon_B$, and 4) strict factorial invariance, in which equal unique factor variances of within- and between-level models are committed. It is given by Θ_{Bg} =Θ_B, and Θ_{Wg}=Θ_W (Kim, Kwok, &Yoon, 2012; Widaman et al., 2010).

Within-level Grouping Variable

Provided that a grouping variable is a level 1 variable, an approach of multiple group multilevel CFA to test multilevel factorial invariance is problematic. It is because the standard procedures of multiple group analysis (dividing individuals into groups according to group membership and conducting multiple group analysis separately but simultaneously) is troublesome given that data dependency cannot be accounted for. For example, individuals in the same clusters are divided into different groups according to

their group membership. In this process, the nested data structures are broken. It is difficult to maintain the features of multiple group analysis and multilevel analysis when the grouping variable is at the within level.

Ryu (2014b) proposed a two-step strategy (multilevel modeling and then multiple group multilevel analysis) to test multilevel factorial invariance. First, decompose the level-1 variables into within-cluster and between-level components, which is given by $y_{ij} = y_{Bj} + y_{Wj}$. Ryu (2014b) pointed out that the decomposition within clusters should be preceded before assigning individuals to groups. In this way, the observation dependency is taken care of before using the multilevel modeling techniques. Second, testing multilevel factorial invariance between groups uses multiple group multilevel analysis techniques. Note that the data dependency has been adjusted through a two-step strategy before imposing constrains for testing different levels of invariance.

Kim, Kwok, and Yoon (2012) adhered to a single-level multiple group analysis approach while making necessary adjustment due to data dependency by adjusting the chi-square statistics and standard error estimates. It can be achieved by using the TYPE = COMPLEX routine in M*plus*. They justified the procedures with two reasons: 1) some researchers (e.g., Muthén & Satorra, 1995; Wu & Kwok, 2012) demonstrated that design-based approach accounted for observation nonindependency in multilevel data, and 2) given a single-level model specification, the multiple group analysis can then be proceeded.

Jak, Oort, and Dolan (2014) proposed the multilevel restricted factor analysis (multilevel RFA) which developed from single-level RFA (Oort, 1992, 1998). In this method, the grouping variables were correlated with the latent factors. Invariance was examined by testing the significance of the effect that the indicator regressed on the grouping variable.

This dissertation focused on the scenario that the grouping variable was at the between level. In the current version of M*plus*, it only has the feature of multiple group analysis of multilevel data when the grouping variable is at the cluster level.

Significance of the Study

The dissertation addressed a pressing issue — level-specific goodness-of-fit indices in evaluating factorial invariance in MSEM in comparable to the standard model fit indices (i.e., overall model fit indices in MSEM). The significance of the study was presented from two aspects as follows.

First, various goodness-of-fit indices to test measurement invariance have been extensively studied in single-level analysis (e.g., Chen, 2007; Cheung & Rensvold, 2002, Meade, Johnson, & Braddy, 2008), but not in MSEM. Given multilevel data are common in educational and psychological research, examining factorial invariance in multilevel setting is in need. Kim, Kwok, and Yoon (2012) studied the impact of a single-level analysis to examine factorial invariance in multilevel data. They found that if ignoring multilevel data structure when examining factorial invariance in multilevel data, increased Type I error was a serious problem. Moreover, Type I error became increasingly serious when ICC and cluster size were larger. Similarly, Pornprasertmanit,

Lee, and Preacher (2014) also warned researchers of potential bias such as low detection of model fit and biased standard errors regardless of in a disaggregated analysis or an aggregated analysis. To sum up, multiple group single-level analysis was not appropriate for examining factorial invariance in multilevel data.

Second, further complicating matters were, various goodness-of-fit indices may behaved differently in response to various degrees of noninvariance in multilevel models (e.g., noninvariance at the between level, the within level, or at both levels). My research question recognized the additional complications in model fit evolution in MSEM. Previous research demonstrated that overall model fit indices provided information of model fit for the entire model but may fail to capture misfits at specific data level, especially at the higher level in MSEM (e.g., Hox, 2010; Hsu, Kwok, Lin, & Acosta, 2015; Ryu & West, 2009; Yuan & Bentler, 2007). For example, chi-square statistics indicated how the model-implied variance covariance structure fitted the observed variance covariance structure in an overall picture, but not within-level model and between-level model specifically. To conclude, overall fit indices didn't reflect the lack of fit for specific level model. A small amount of noninvariance which occurs in the between-level model maybe overlay by the dominant invariance in the within-level model. Therefore, noninvariance at the between level may be more difficult to be detected and easily missed out. Any of these cases suggested a further need of examining factorial invariance at specific data level in MSEM. Yuan and Bentler (2007) and Ryu and West (2009) proposed model fit evaluation for each level of model separately in MSEM. To date, no study has yet explored their approaches for detecting model

misspecification due to factorial noninvariance and was thus the focus of the current study.

In this study, the sensitivity of goodness-of-fit indices to lack of invariance were systematically studied using multiple group multilevel CFA. The performance of level-specific model fit evaluation was also compared to the standard model fit evaluation. This study should be viewed a step forward in research of measurement invariance testing which extended to MSEM framework.

METHOD

A Monte Carlo study was conducted using M*plus* 7.3 (Muthén & Muthén, 1998-2012) and a free R package (M*plus* automation, Hallquist & Wiley, 2013) to examine the level-specific model fit statistics for testing factorial invariance in multilevel data. Meanwhile, the level-specific model fit evaluation was compared to the standard approach. The multiple group multilevel modeling features in M*plus* were used to generalize and analyze data. As described previously, this feature is only available when the grouping variable is at the cluster level. The experimental conditions that were manipulated in this study were justified from two perspectives: 1) factors that might affect the sensitivity of goodness-of-fit indices for factorial invariance testing in MSEM, 2) factors that are characteristics of multilevel data structure (e.g., number of clusters, cluster sizes, and interclass correlation).

Previous simulation studies on measurement invariance testing, studies on multilevel modeling, and studies on fit indices in MSEM were reviewed (e.g., Chen,

2007; Cheung & Rensvold, 2002; Hox & Mass, 2001; Hsu, Kwok, Lin, & Acosta, 2015; Kim, Kwok, & Yoon, 2012; Maas & Hox, 2005; Meade, Johnson, & Braddy, 2008; Millsap & Kwok, 2004; Ryu & West, 2009). The population model, study conditions, and data generation procedures, were detailed as follows.

Population Model

One of the foci of this simulation study was to examine the sensitivity of goodness of fit indices (i.e., chi-square, CFI, TLI, RMSEA, and SRMR) for testing factorial invariance in multiple group multilevel CFA. For this purpose, data were generated based on a two-level factor model with a between-level grouping variable which was presented in Figure 4.1 for illustration. As shown in Figure 4.1, eight indicators loaded on one latent factor in the within- and between-level models, respectively.

Especially following Kim, Kwok, and Yoon's (2012) study, the population parameters in their studies were adopted in this study. The factor loadings for the between-level model (B) and the within-level model (W) were given as:

$$\Lambda_B = \Lambda_W = \begin{bmatrix} 0.7 \\ 0.9 \\ 0.6 \\ 0.6 \\ 0.8 \\ 0.6 \\ 0.9 \\ 0.7 \end{bmatrix}.$$

The between-level and within-level model had the same factor structure (eight indicators per one factor). In addition, the between-level factor loadings equaled the within-level factor loadings. The factor loadings were higher than those in Kim, Kwok, and Yoon's

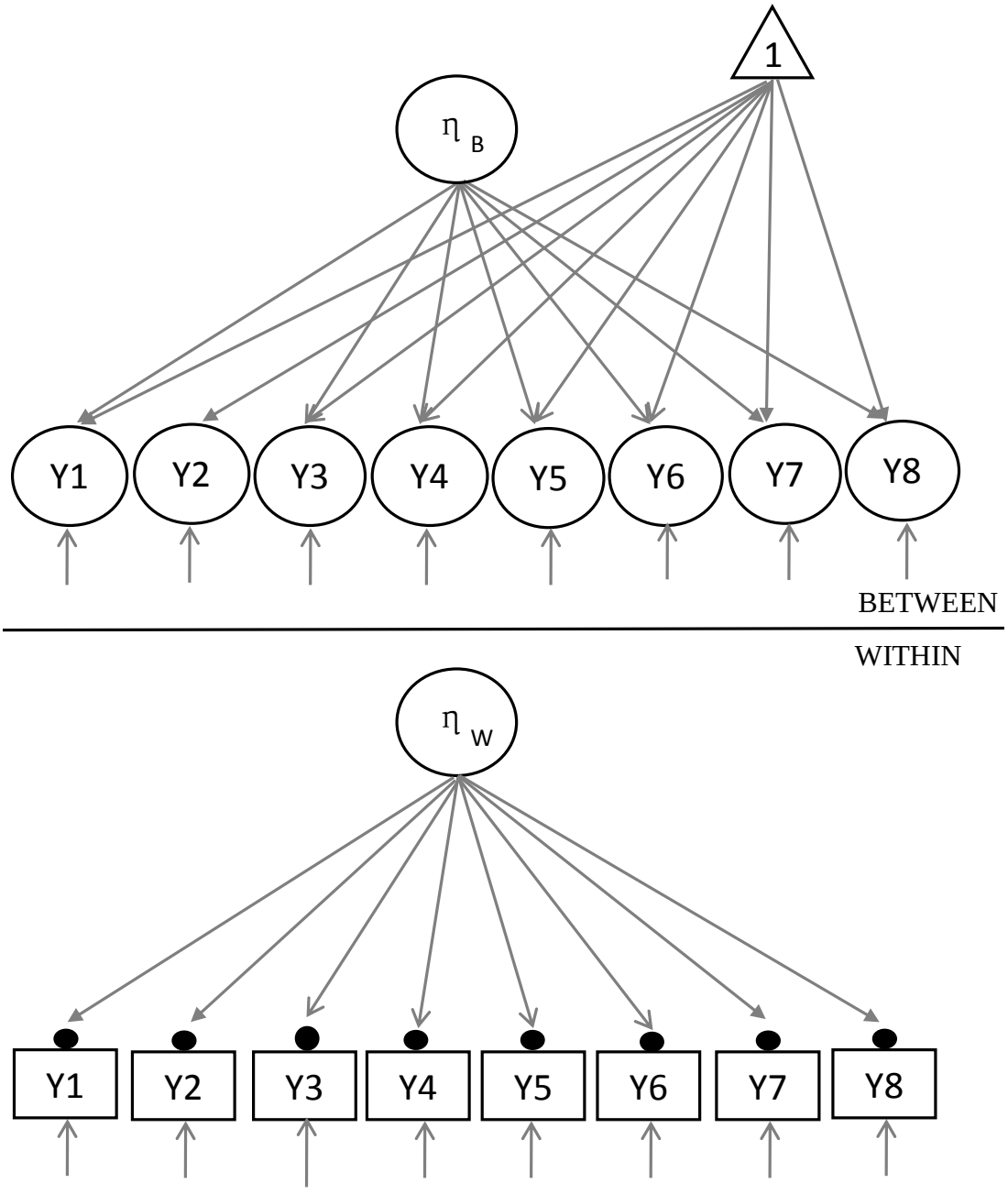

FIGURE 2 Two-level Measurement Model with One Factor and Eight Indicators

(2012) study. They ranged from 0.6 to 0.9 in this study while ranged from 0.3 to 0.9 in Kim, Kwok, and Yoon's study. It was because that in this study two levels of between-group differences (small and large) were manipulated for different numbers of indicators. In Kim, Kwok, and Yoon's (2012) study, the between-group difference was held constant for one factor loading between two groups, and was set to 0.5.

The other population parameter values were as follows. The within-level factor variance was 1.0. The between-level factor variances varied to create different levels of ICCs, which were detailed in a later section. The between-level factor mean was set to zero. The between-level intercepts were set to zero. When evaluating strict invariance (i.e., invariance in intercepts), the focal group had different intercept values (not zero), which was discussed in the next section. No within-level factor mean and intercepts are specified in the population model (see Heck, 2001; Muthén & Muthén, 1998); and 4) the residual variances of eight indicators equaled 0.25 at both the within level and the between level (Hox & Mass, 2001). That is,

$$\Theta_B = \Theta_W = \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.25 \end{bmatrix},$$

which showed that the indicator unique variances were uncorrelated with each other (random measurement error assumption).

The focal group and the reference group had an identical two-level model structure. Nevertheless, to examine the sensitivity of goodness-of-fit indices to lack of

factorial invariance, two scenarios were created. In scenario 1, noninvariance at between-level factor loading was manipulated. Specifically, the between-level factor loadings in the focal group were higher than that for the reference group. In scenario 2, noninvariance at between-level intercepts was manipulated. That is, the focal group had higher between-level intercepts than the reference group. Meanwhile, the factor loadings were kept invariant across two groups. Taken together, metric invariance (equal factor loadings) and scalar invariance (equal intercepts) were tested respectively in the two scenarios.

Study Conditions

In this study, total six experimental variables that may affect factorial invariance testing in multilevel SEMs were considered. They included two levels of location of noninvariance, five proportions of invariance, two levels of between-group differences, four levels of number of clusters, two levels of cluster size, and three levels of ICC. In all, the study design was a factorial design with $2 \times [5 \times 2 \times 4 \times 2 \times 3 - 24$ (noninvariant conditions)] = $2 \times (240 - 24) = 384$ cells. For each cell, 1,000 replications were generated, resulting in a total of 384,000 replications for all conditions. In the following sections, each experimental condition was presented with details.

Location of Noninvariance. Two kinds of noninvariance have been commonly tested in empirical studies. One was noninvariance in factor loadings (metric invariance). The other was noninvariance in intercepts (scalar invariance). As discussed earlier, the purpose of this study was to examine the sensitivity of the goodness-of-fit indices for testing invariance in factor loadings and intercepts. Therefore, noninvariant factor

loadings and intercepts in the between-level model were generated along with invariant conditions for comparison.

Number of Groups. Two groups were chosen in accord with other simulation studies on measurement invariance testing (e.g., Asparouhov & Muthén, 2012; Chen, 2007; Cheung & Rensvold, 2002; Kim, Kwok, & Yoon, 2012).

Proportion of Noninvariance. In the two-level model, while the within-level model was held invariant across groups, different proportions of item noninvariance were simulated in the between-level model across groups. Chen (2007) investigated five conditions of noninvariance (i.e., 0%, 25%, 50%, 70%, or 100%) in factor loadings and intercepts. The sensitivity of goodness of fit indices was investigated in response to lack of invariance. French and Finch (2008) examined three conditions of noninvariance: a comparison condition (0% of noninvariance), a low contamination condition (17% of noninvariance), and a high contamination condition (34% of noninvariance). In this study, five conditions (i.e., 0%, 25%, 50%, 70%, or 100%) of lack of invariance were replicated in the between-level factor loadings and intercepts. The 100% invariance condition served as a comparison with all factor loadings or intercepts invariant across groups. The difference in factor loadings and intercepts conditions was discussed below with more details.

Levels of Invariance. As presented above, compared to the reference group the focal group had different factor loadings or intercepts in the between-level model. In scenario 1, in which factor loading invariance (metric invariance) was examined, the two groups differed in between-level factor loadings. For example, in the 50% invariance

conditions the reference group had higher factor loadings compared to the focal group as follows:

$$\text{Reference Group: } \Lambda_B = \Lambda_W = \begin{bmatrix} 0.7 \\ 0.9 \\ 0.6 \\ 0.6 \\ 0.8 \\ 0.6 \\ 0.9 \\ 0.7 \end{bmatrix}, \qquad \text{Focal Group: } \Lambda_W = \begin{bmatrix} 0.7 \\ 0.9 \\ 0.6 \\ 0.6 \\ 0.8 \\ 0.6 \\ 0.9 \\ 0.7 \end{bmatrix}, \ \Lambda_B = \begin{bmatrix} 0.7 \\ 0.9 \\ 0.6 \\ 0.6 \\ 0.3 \\ 0.1 \\ 0.4 \\ 0.2 \end{bmatrix}.$$

In this example, note that the reference group had the same factor loading for both the with-level and the between-level model. The within-level model was invariant between the focal and reference group. For the focal group, 50% factor loadings in the between-level model were noninvariant with 0.5 between-group difference in factor loadings relative to the reference group. In scenario 2, difference in the between-level intercepts were manipulated. For example, for the 50% invariant intercept conditions, the reference group and the focal group differed in half of the total intercepts with 0.2 between-group difference, given as:

$$\text{Reference Group: } \nu_B = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}, \qquad \text{Focal Group: } \nu_B = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix},$$

in which all intercepts were 0.2 for the reference group while for the focal group four between-level intercepts were zero (50% noninvariant intercepts).

Between-group Differences. When the reference and focal group differed in factor loading, two magnitudes of between-group difference were simulated. When between-group difference in between-level factor loadings was 05, it was considered as a large effect size (Kim, Kwok, & Yoon, 2012). The other was 0.2 which was considered as small between-group difference and represented a comparison (French and Finch, 2008). Similarly, a difference of 0.2 or 0.5 in the between-level intercepts was simulated for the noninvariant intercepts (e.g., the intercept for reference group was set to 0.2, whereas intercept for the focal group was 0). Chen (2007) examined two patterns of invariance, including mixed and uniform, which only the uniform pattern was considered in this study. That is, the focal group always had higher factor loadings or intercepts than the reference group.

Number of Clusters and Cluster Sizes. Following Kim, Kwok, & Yoon's (2012) study, four conditions of the number of clusters (CN) were chosen (i.e., 30, 50, 80, and 100). Two cluster sizes (CS) included 10 and 20.

Intra-Class Correlation (ICC). ICC in this study meant the intraclass correlation (ICC) for the latent factor (c.f., Heck & Thomas, 2009). Three levels of ICC were manipulated in this study, which were consistent to some simulation studies in MSEM (e.g., Hox & Mass, 2001; Kim, Kwok, & Yoon, 2012; Wu & Kwok, 2012). The within-level factor variance was fixed at 1. Three levels of latent factor ICC were created by varying the between-level factor variance. In specific, the high ICC (i.e., $\rho = 0.5$), the medium ICC (i.e., $\rho = 0.2$) and the small ICC (i.e., $\rho = 0.1$) conditions had the between-level factor variance set to 1, 0.25, and 0.11, respectively.

Data Analysis

All simulated data sets were analyzed in M*plus* 7.11 (Muthén & Muthén, 1998-2012) with maximum likelihood estimation with robust standard errors (MLR). Invariance in two locations were examined including factor loadings and intercepts. First, factor loading invariance was examined, followed by test of intercept invariance.

In model 1 (M1), metric invariance models were specified with equal between-level factor loadings across two groups while the within-level models were saturated. In contrast, in model 2 (M2), the constrained equal between-level factor loadings were released to be freely estimated in the focal group. Meanwhile, the within-level models were still saturated. In this way, level-specific model fit indices were obtained for between-level models of M1 and M2 when the within-level models were saturated. Model 3 (M3) and model 4 (M4) were specified in the similar way. However, both in M3 and M4, the within-level models were not saturated but specified identically to the between-level model. In model 3 (M3), equal between-level factor loadings constraints were imposed. In model 4 (M4), equal between-level factor loadings were set to be freely estimated in the focal group. Therefore, standard model fit indices (overall model fit) were obtained for M3 and M4.

When invariance in intercepts was examined, scalar invariance models had equal between-level intercepts across two groups while the within-level models were saturated. As a comparison, the model constraints were then freely estimated in the focal group while within-level models were still saturated. In this way, level-specific model fit indices were obtained for between-level models when the within-level models were

saturated. Standard model fit indices were obtained when equal between-level intercepts were imposed but the within-level models were not saturated. Next, the constrained equal between-level intercepts were freely estimated in the focal group. Again, the within-level models were not saturated. In this way, standard model fit indices (for the entire model) were derived.

The results for testing metric invariance and scalar invariance were presented separately (in two sections). First, means and standard deviations of level-specific model fit indices (i.e., X_B^2, CFI_B, $RMSEA_B$, TLI_B, and $SRMR_B$) and standard model fit indices (i.e., overall model fit indices including X^2, CFI, RMSEA, TLI, and $SRMR_B$) were summarized. Second, a separate five-way analysis of variance (ANOVA) was performed for each index to assess the effects of those study variables on changes of goodness-of-fit indices. Last, the means and standard deviations of the ΔGFIs were summarized. For the level-specific approach, changes in five indices including ΔX_B^2, ΔCFI_B, $\Delta RMSEA_B$, ΔTLI_B, and $\Delta SRMR_B$ were listed. For the standard approach, the changes in five overall indices including ΔX^2, ΔCFI, ΔRMSEA, ΔTLI, and $\Delta SRMR_B$ were presented.

RESULTS

In the result section (either text or tables), results for the level-specific model fit evaluation approach and the standard model fit approach were reported separately. As described earlier, level-specific model fit evaluation approach produced within-level and between-level model fit indices separately. Given only between-level noninvariance was manipulated and examined in this study, model fits only for the between-level model

were reported. On the other hand, within-level models were saturated, resulting in perfect model fit. Therefore, results for the within-level models were not included in the result summary. For the standard model fit approach, $SRMR_W$ for the within-level model and $SRMR_B$ for the between-level model were both available in M*plus*. To be compared to the level-specific model evaluation, $SRMR_B$ for the between-level model was reported in the standard approach and $SRMR_W$ were not included. The other model fit indices (X^2, CFI, RMSEA, and TLI) had general values for the entire model. In the following sections, results of testing metric invariance were presented first, following by results of testing scalar invariance.

Testing Metric Invariance

Nonconvergence Problems. Overall, the level-specific fit evaluation approach had similar pattern in terms of model convergence compared to the standard model fit evaluation. Small variation in convergence rate was observed across different proportions of invariance conditions (i.e., 0% to 100%). In addition, different between-group difference (i.e., 0.2 and 0.5) had little impact on the rate of convergence. For simplicity, detailed convergence rates were not reported for these two design conditions.

As shown in Table 4.2, for the small, medium, and large ICC levels the average convergent solutions were 99.02%, 99.47%, and 99.81%, respectively. For the large number of clusters conditions (e.g., 80 or 160), almost all models converged properly. More nonconverged solutions occurred when the number of clusters was small (e.g., 30) and ICC was also small (e.g., 0.1). To obtain a higher rate of convergent solutions, a large number of clusters was needed for testing metric invariance in MSEM, especially

when ICC was small. For testing metric invariance in the LR test, convergence problems were mainly issues in the unconstrained models. In other words, when equal constraints across groups were released (all equality constraints were freely estimated across groups), more convergence problems occurred. Given small percentages of nonconvergence out of all replications, nonconverged solutions were screened out and not included in the result summaries. In addition, a few replications (e.g., one or two out of 1,000 replications when the number of cluster was 30) with extreme chi-square statistics (e.g., above 2,000) and with negative TLI values were also deleted.

Improper Solutions. The majority *Mplus* warning messages in some replications indicated theta problems (negative residual variance estimates). Overall, ICC, number of clusters, and cluster sizes had impact on the average rate of problematic replications. Table 4.3 presented the average rates of problematic replications in terms of the three factors.

In total, 11.87% problematic replications (negative residual variance estimates) occurred. Only 1.48% problematic solutions were encountered in large ICC (e.g., 0.5 in this study) conditions. The rates were 8.96% and 25.16% in medium ICC conditions (e.g., 0.2 in this study) and small ICC conditions (e.g., 0.1 in this study), respectively. The number of problematic solutions decreased with increasing sample sizes (cluster size) as well as with increasing number of clusters. The results showed that higher ICC along with larger number of clusters and more observations per cluster resulted in less errors (i.e., negative residual variance).

A further check with those problematic replications was conducted. It was found

that all negative residual variance estimates in those problematic replications were not statistically significantly different from zero. On the other hand, when the number of replications increased to 2,000 for conditions with problematic replications, no substantial changes in model fit indices were found. Therefore, the original results with 1,000 replications were reported which included replications with negative residual variance estimates. In other words, replications exhibiting error messages of negative residual variance were not treated as improper solutions in this study. In the section of testing scalar invariance, replications with negative residual variance estimates are not excluded either, which was discussed in a later section.

Model Fit Indices. Results of the average GFIs (i.e., CFI, RMSEA, SRMR, and TLI) as well as chi-square statistics for testing between-level metric invariance were summarized in Table 4.4 along with standard deviations. Following Hu and Bentler's (1999) recommendations, the following cutoff criteria for model fit were used, including: a) CFI equal or larger than .95, b) RMSEA equal or smaller than .08, c) SRMR equal or smaller than .08, and d) TLI equal or larger than .95.

As shown in Table 4.4, in the invariant conditions (100% invariance), the mean X^2 were 55.52 (SD = 13.23) and 96.04 (SD = 15.88) for the level-specific and standard approach, respectively. This evidence suggested between-level metric invariance as hypothesized (for X^2 with df of 48 and 88, the critical value is 65.17 and 110.90, respectively). In the noninvariant conditions (0% to 75% invariance), means of ${X^2}_{\mathrm{B}}$ were all above the critical value for the level-specific approach, indicating lack of invariance as hypothesized. However, it was noted that ${X^2}_{\mathrm{B}}$ was not stable in response to

TABLE 4.2 Summary of Convergence Rates for Factor Loading Invariance Testing

		Level-specific approach		Standard approach	
	CN/CS	M1	M2	M3	M4
Small ICC	30/10	97.24%	93.29%	96.23%	91.83%
	30/20	99.60%	97.65%	99.64%	97.75%
	50/10	99.80%	98.63%	99.71%	98.35%
	50/20	99.98%	99.92%	100%	99.86%
	80/10	99.97%	99.77%	100%	99.82%
	80/20	100%	99.99%	100%	99.99%
	160/10	99.99%	99.80%	100%	99.83%
	160/20	100%	100%	100%	100%
Average					99.02%
Medium ICC	30/10	98.51%	95.06%	98.06%	94.87%
	30/20	99.91%	99.03%	99.95%	99.11%
	50/10	100%	99.43%	99.96%	99.45%
	50/20	100%	99.94%	100%	99.95%
	80/10	100%	99.96%	100%	99.95%
	80/20	100%	100%	100%	100%
	160/10	100%	99.96%	100%	99.95%
	160/20	100%	100%	100%	99.99%
Average					99.47%
Large ICC	30/10	99.70%	98.24%	99.19%	98.33%
	30/20	100%	99.69%	100%	99.55%
	50/10	100%	99.64%	100%	99.69%
	50/20	100%	99.95%	100%	99.95%
	80/10	100%	99.94%	100%	99.99%
	80/20	100%	100%	100%	100%
	160/10	100%	99.94%	100%	99.99%
	160/20	100%	100%	100%	100%
Average					99.81%

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
M1 & M3: Metric invariance model
M2 & M4: Metric invariance unconstrained model

TABLE 4.3 Summary of Percentage of Problematic Replications for Factor Loading Invariance Testing

		Level-specific approach		Standard approach	
	CN/CS	M1	M2	M3	M4
Small ICC	30/10	35.14%	67.67%	34.61%	66.51%
	30/20	23.36%	62.68%	23.19%	62.39%
	50/10	16.47%	54.29%	16.30%	53.97%
	50/20	8.26%	41.68%	8.18%	41.63%
	80/10	7.27%	36.90%	7.28%	36.90%
	80/20	1.62%	23.76%	1.61%	23.73%
	160/10	0.89%	14.14%	0.86%	14.28%
	160/20	0.02%	9.82%	0.02%	9.81%
Average					25.16%
Medium ICC	30/10	8.69%	39.38%	8.67%	39.40%
	30/20	2.64%	26.81%	2.68%	32.34%
	50/10	1.10%	20.17%	1.01%	20.20%
	50/20	0.19%	13.80%	0.18%	13.64%
	80/10	0.12%	10.21%	0.10%	10.08%
	80/20	0.00%	8.40%	0.01%	8.40%
	160/10	0.00%	5.11%	0.00%	5.09%
	160/20	0.00%	4.13%	0.00%	4.12%
Average					8.96%
Large ICC	30/10	0.17%	8.68%	0.08%	8.74%
	30/20	0.06%	6.59%	0.01%	6.42%
	50/10	0.00%	4.49%	0.00%	4.61%
	50/20	0.00%	2.94%	0.00%	2.92%
	80/10	0.00%	0.00%	0.00%	0.00%
	80/20	0.00%	0.83%	0.00%	0.86%
	160/10	0.00%	0.00%	0.00%	0.00%
	160/20	0.00%	0.04%	0.00%	0.03%
Average					1.48%

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
M1 & M3: Metric invariance model
M2 & M4: Metric invariance unconstrained model

different degree of invariance. The relation between the degree of invariance and chi-square statistics was nonmonotonic. With respect to the standard approach, when 75% items factor loadings were invariant (25% factor loadings were noninvariant), X^2 (i.e., M = 108.25, SD = 26.80) failed to reject the null hypothesis of invariance. As the invariance proportion increased from 0% to 50%, means of X^2 were all above the critical value, indicating lack of invariance. However, the relation between the degree of invariance and chi-square statistics was nonmonotonic in the standard approach.

On average, the CFI values ranged from 0.997 to 0.999 with small SDs across different proportions of invariance (i.e., 0% to 100% invariance), indicating good model fit according to the suggested cutoff. Similarly, TLI ranged from 0.993 to 0.997 and from 0.996 to 0.998 for the level-specific and standard approach, respectively. Large CFI and TLI values (> 0.99) and negligible changes across different proportions of invariance suggested good model fit whereas some factor loadings were noninvariant across groups, either in the level-specific approach or in the standard approach. In conclusion, CFI and TLI were not sensitive indices for examining between-level factor loading invariance according to suggested cutoff criteria.

In the invariant conditions (100% invariance), RMSEA had smaller values than the suggested cutoff value of 0.08 regardless of level-specific approach or standard approach, which indicated metric invariance as expected. In the noninvariant conditions (0% to 75% invariance), means of RMSEA were smaller than the cutoff of 0.08 as well, indicating metric invariance. RMSEA decreased in a small degree as the degree of invariance increased. In conclusion, whether in the level-specific approach or standard

approach, RMSEA was unable to detect lack of invariance based on the traditional cutoff value of 0.08.

The intriguing finding for testing metric invariance was $SRMR_B$ in the level-specific approach had identical results to $SRMR_B$ in the standard approach. The major reason was that $SRMR_B$ in the standard approach was also an index for the between-level model only (M*plus* provides SRMB for the between level model and SRMR for the within level mode separately). The results further showed that level-specific model fit index was promising for examining invariance in MSEM, which allowed for examination of invariance in models in models at specific data level. More importantly, the change of SRMR exhibited a decreasing trend when the degree of invariance increased. In other words, the relation between the degree of invariance (0% to 100% invariance) and SRMR values was monotonic. In addition, SRMR was sensitive to detect lack of metric invariance in the between-level models. For example, in the invariant conditions (100% invariance), the mean SRMR was 0.086 (SD = 0.042), which suggested marginally good model fit. In the noninvariant conditions (0% to 75% invariance), means of SRMR were all larger than 0.08, indicating noninvariant factor loadings.

To determine factors that affected changes in fit statistics, a five-way analysis of variance (ANOVA) was conducted for each fit index separately. Specifically, the effects of different proportion of invariance (INV), intraclass correlation (ICC), number of clusters (CN), cluster size (CS), and between-group difference (DIF) on the changes in five target fit indices were examined. The ANOVA results were presented in Table 4.5

and Table 4.6 for the level-specific approach and the standard approach, respectively. Eta-squared effects sizes (η^2) were calculated, which were given as the ratio of the type III sum of squares of each factor or interaction ($SS_{between}$) to the corrected total sum of squares (SS_{total} without the intercept proportion). Eta-squared effects sizes (η^2) reflected the percentages of total variance explained by the above five factors and their interactions.

As shown in Table 4.5 for the level-specific approach, a good proportion of variance in all ΔGFIs (i.e., ΔCFI_B, $\Delta RMSEA_B$, $\Delta SRMR_B$, and ΔTLI_B) along with $\Delta\chi_B^2$ was explained by between-group difference. In other words, the magnitude of between-group difference had the largest effect on the changes of fit indices. Larger between-group difference predicted greater changes in GFIs and χ_B^2. ICC level was the second factor that affected changes in all fit indices with a large effect. The interaction of ICC and between-group difference had the third largest effect. As shown in Table 4.6, similar patterns were observed for the standard approach. To sum up, the changes of fit indices were affected substantially by between-group difference, ICC, and their interactions in both level-specific approach and the standard approach. The changes were less affected by the other factors including number of clusters, cluster size, and proportion of invariance.

Fit Indices Sensitivity in Metric Invariance Tests. As discussed previously, changes in fit indices were used to test metric invariance hypotheses, which were the differences between the fit of constrained models and unconstrained models. ΔGFIs (i.e.,

TABLE 4.4 Means and SDs (in parenthesis) of Model Fit Indices for Metric Invariance Models

Invariance (%)	Level-specific Approach [a]					"Standard" Approach				
	$\chi^2(48)_B$	CFI_B	$RMSEA_B$	$SRMR_B$	TLI_B	$\chi^2(88)$	CFI	RMSEA	$SRMR_B$	TLI
0%	69.62	0.998	0.019	0.152	0.995	109.84	0.998	0.015	0.152	0.997
	(25.94)	(0.003)	(0.014)	(0.068)	(0.007)	(27.15)	(0.003)	(0.011)	(0.068)	(0.004)
25%	74.25	0.997	0.021	0.136	0.994	114.43	0.997	0.016	0.136	0.997
	(35.42)	(0.003)	(0.015)	(0.054)	(0.008)	(35.65)	(0.004)	(0.012)	(0.054)	(0.006)
50%	77.44	0.997	0.022	0.127	0.993	117.38	0.997	0.016	0.127	0.996
	(41.90)	(0.004)	(0.016)	(0.051)	(0.009)	(41.41)	(0.004)	(0.012)	(0.051)	(0.005)
75%	67.97	0.998	0.019	0.105	0.995	108.25	0.998	0.014	0.105	0.997
	(25.83)	(0.003)	(0.014)	(0.041)	(0.007)	(26.80)	(0.003)	(0.011)	(0.041)	(0.004)
100%	55.52	0.999	0.013	0.086	0.997	96.04	0.999	0.010	0.086	0.998
	(13.23)	(0.002)	(0.013)	(0.042)	(0.005)	(15.88)	(0.002)	(0.010)	(0.042)	(0.003)

Note. [a] In the level-specific approach, the within-level model is saturated.

B denotes between-level.

CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean square residual; TLI = the Tucker-Lewis index.

TABLE 4.5 Eta-Squared Effect Size Estimates for Study Variables on the Change in Fit Indices (Level-specific Indices)

Variables	Level-specific Indices				
	$\Delta\chi^2(8)_B$	ΔCFI_B	$\Delta RMSEA_B$	$\Delta SRMR_B$	ΔTLI_B
INV	0.97%	0.51%	0.75%	6.47%	0.42%
ICC	11.26%	5.54%	10.65%	13.38%	4.67%
CN	3.75%	0.10%	0.60%	0.02%	0.04%
CS	0.10%	0.38%	0.11%	0.00%	0.18%
DIF	11.08%	5.65%	12.36%	20.00%	4.92%
INV * ICC	1.61%	0.84%	1.08%	3.54%	0.71%
INV * CN	0.32%	0.02%	0.03%	0.12%	0.02%
INV * CS	0.02%	0.03%	0.01%	0.00%	0.02%
INV * DIF	0.52%	0.29%	0.24%	2.21%	0.24%
ICC * CN	4.11%	0.10%	1.01%	0.14%	0.16%
ICC * CS	0.05%	0.41%	0.20%	0.00%	0.37%
ICC * DIF	9.04%	4.61%	5.70%	12.71%	3.55%
CN * CS	0.06%	0.02%	0.00%	0.00%	0.00%
CN * DIF	3.49%	0.02%	0.17%	0.30%	0.02%
CS * DIF	0.10%	0.34%	0.10%	0.01%	0.31%
INV * ICC * CN	0.50%	0.02%	0.04%	0.03%	0.02%
INV * ICC * CS	0.01%	0.07%	0.02%	0.00%	0.06%
INV * ICC * DIF	0.79%	0.41%	0.24%	1.31%	0.33%
INV * CN * CS	0.00%	0.00%	0.00%	0.00%	0.00%
INV * CN * DIF	0.15%	0.01%	0.00%	0.02%	0.01%
INV * CS * DIF	0.01%	0.02%	0.00%	0.00%	0.01%
ICC * CN * CS	0.01%	0.04%	0.09%	0.01%	0.05%
ICC * CN * DIF	3.50%	0.01%	0.04%	0.06%	0.01%
ICC * CS * DIF	0.04%	0.36%	0.14%	0.00%	0.28%

(continued)

TABLE 4.5 (continued)

Variables	Level-specific Indices				
	$\Delta\chi^2(8)_B$	ΔCFI_B	$\Delta RMSEA_B$	$\Delta SRMR_B$	ΔTLI_B
CN * CS * DIF	0.04%	0.00%	0.01%	0.00%	0.00%
INV * ICC * CN * CS	0.01%	0.01%	0.00%	0.00%	0.01%
INV * ICC * CN * DIF	0.26%	0.01%	0.02%	0.00%	0.01%
INV * ICC * CS * DIF	0.00%	0.03%	0.00%	0.00%	0.02%
INV * CN * CS * DIF	0.00%	0.00%	0.00%	0.01%	0.00%
ICC * CN * CS * DIF	0.02%	0.00%	0.00%	0.00%	0.00%
INV * ICC * CN * CS * DIF	0.00%	0.00%	0.00%	0.00%	0.00%

Note: INV = Invariance proportion. ICC = Intra-class correlation. CN = Number of clusters. CS = Cluster size. DIF = Between-group difference.

$_B$ denotes between-level model.

CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean square residual; TLI = the Tucker-Lewis index.

TABLE 4.6 Eta-Squared Effect Size Estimates for Study Variables on the Change in Fit Indices (Standard Indices)

Variables	Standard Indices				
	$\Delta\chi^2(8)$	ΔCFI	ΔRMSEA	$\Delta SRMR_B$	ΔTLI
INV	1.06%	0.56%	0.84%	6.46%	0.54%
ICC	12.06%	5.82%	11.78%	13.39%	5.54%
CN	4.04%	0.24%	0.36%	0.02%	0.06%
CS	0.11%	0.53%	0.18%	0.00%	0.46%
DIF	12.29%	6.37%	13.69%	20.04%	6.45%
INV * ICC	1.72%	0.92%	1.16%	3.54%	0.89%
INV * CN	0.37%	0.03%	0.03%	0.12%	0.04%
INV * CS	0.02%	0.03%	0.01%	0.00%	0.03%
INV * DIF	0.55%	0.34%	0.28%	2.21%	0.30%
ICC * CN	4.85%	0.10%	1.03%	0.13%	0.16%
ICC * CS	0.06%	0.42%	0.19%	0.00%	0.41%
ICC * DIF	10.13%	5.43%	7.31%	12.71%	4.81%
CN * CS	0.07%	0.06%	0.02%	0.00%	0.02%
CN * DIF	4.01%	0.02%	0.24%	0.30%	0.02%
CS * DIF	0.11%	0.41%	0.10%	0.01%	0.43%
INV * ICC * CN	0.56%	0.02%	0.04%	0.03%	0.03%
INV * ICC * CS	0.01%	0.07%	0.02%	0.00%	0.07%
INV * ICC * DIF	0.83%	0.45%	0.27%	1.31%	0.40%
INV * CN * CS	0.01%	0.01%	0.00%	0.00%	0.01%
INV * CN * DIF	0.17%	0.01%	0.01%	0.02%	0.01%
INV * CS * DIF	0.01%	0.03%	0.01%	0.00%	0.02%
ICC * CN * CS	0.01%	0.03%	0.08%	0.01%	0.04%
ICC * CN * DIF	3.95%	0.01%	0.08%	0.06%	0.01%
ICC * CS * DIF	0.04%	0.42%	0.17%	0.00%	0.37%

(continued)

TABLE 4.6 (continued)

Variables	Standard Indices				
	$\Delta\chi^2(8)$	ΔCFI	ΔRMSEA	$\Delta SRMR_B$	ΔTLI
CN * CS * DIF	0.04%	0.00%	0.01%	0.00%	0.00%
INV * ICC * CN * CS	0.01%	0.01%	0.00%	0.00%	0.01%
INV * ICC * CN * DIF	0.29%	0.01%	0.02%	0.00%	0.01%
INV * ICC * CS * DIF	0.01%	0.03%	0.00%	0.00%	0.03%
INV * CN * CS * DIF	0.01%	0.00%	0.00%	0.00%	0.00%
ICC * CN * CS * DIF	0.02%	0.00%	0.00%	0.00%	0.00%
INV * ICC * CN * CS * DIF	0.00%	0.01%	0.00%	0.00%	0.01%

Note: INV = Invariance proportion. ICC = Intra-class correlation. CN = Number of clusters. CS = Cluster size. DIF = Between-group difference.

$_B$ denotes between-level model.

CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean square residual; TLI = the Tucker-Lewis index.

ΔCFI, ΔRMSEA, ΔSRMR, and ΔTLI) along with $\Delta\chi^2$were summarized in Table 4.7 to Table 4.14 in terms of different number of clusters and cluster sizes. Specifically, the effects of ICC, number of clusters, cluster size, proportion of factor loading invariance, and between-group difference on the change of model fit statistics were examined. In the next, changes in each fit index were discussed.

ΔX^2. Small difference was observed comparing the level-specific approach with the standard approach. As factors identified in ANOVA, changes in ΔX^2 appeared overall to depend on ICC, between-group difference, and number of clusters. The effects of cluster size and proportion of invariance were not substantial. First, ΔX^2 tended to be larger in large ICC conditions relative to small ICC conditions. Second, the ΔX^2 values increased as the magnitude of the between-group difference increased. Third, with larger number of clusters, ΔX^2 became larger. In contrast, the changes due to different cluster size were small, which meant that the number of cluster has larger effect than the cluster size on changes of comparing the constrained models to the unconstrained models. Last, the changes of X^2 showed no consistent pattern due to different proportions of invariance regardless of level-specific approach or standard approach. For example, in large ICC conditions, ΔX^2 showed the least changes when 75% items were invariant (75% invariance), whereas ΔX^2 showed the largest changes when 50% items were invariant (50% invariance).

In invariant conditions (i.e., 100% invariance), all ΔX^2 were smaller than the critical value (i.e., χ^2_{critical} (0.05, 8) = 15.50), which indicated metric invariance. This evidence suggested that chi-square difference test performed well when all factor

loadings were invariant. In the noninvariant conditions (i.e., 0% to 75% invariance), the performance of ΔX^2 was not stable. Only in large ICC (e.g., 0.5) and large between-group difference (e.g., 0.5) conditions, $\Delta\chi^2$ was larger than the critical value, regardless of the level-specific approach or the standard approach. The evidence showed that only in those settings, the chi-square difference test would be successful in detecting noninvariant factor loadings. However, in small ICC (e.g., 0.1) and small between-group difference (e.g., 0.2) conditions, the chi-square difference test may fail to detect noninvariant factor loadings, irrespective of the proportion of invariance (0% invariance to 75% invariance) (see Table 4.7 to Table 4.14). In medium ICC and small between-group difference conditions, similar patterns were observed. To sum up, ICC had a substantial impact on the sensitivity of $\Delta\chi^2$ for examining invariance in the between-level factor loadings. When between-group difference was small, large ICC was needed to detect lack of factor loading invariance in the between-level model. In addition, both the number of clusters and cluster size should be large (e.g., CN = 80 and CS = 20). When between-group difference was large, large ICC was a sufficient condition to detect noninvariant between-level factor loadings regardless of the number of clusters and cluster size.

ΔCFI. The ΔCFIs values in the level-specific approach and in the standard approach were identical to each other. Consistent with findings of ANOVA, two factors (i.e., ICC and between-group difference) affects the changes in CFIs. The effects of other factors (e.g., proportion of invariance, number of cluster, and cluster size) were not obvious. Specifically, only in large ICC and large between-group difference conditions,

ΔCFI was larger than -0.002. Other than that, ΔCFI approximated zero (e.g., in small ICC and small between-group difference conditions).

In the invariant conditions, ΔCFIs were generally around 0.000, indicating metric invariance. In the noninvariant conditions, changes in CFI were not salient. For example, when ICC was small (0.1 in this study) and medium (0.2 in this study), ΔCFI was the smallest with a small range from -0.002 to 0.000. When ICC was large (0.5 in this study) and between-group difference was large (0.5 in this study), changes in CFI slightly increased. As is shown in Table 4.13, CFI varied from -0.009 to -0.005.

In conclusion, ΔCFI was not a sensitive index for examining invariance in factor loadings in MSEM, especially when ICC and between-group difference were small. ΔCFI tended to be larger when ICC and between-group difference increased. However, the pattern of changes was not consistent (from 0% to 75% invariance). For example, in large ICC conditions, when 50% of items were invariant, the changes were the largest, whereas the changes were the smallest when 75% of items were invariant. In summary, the sensitivity of ΔCFI to lack of between-level factor loading invariance was affected by ICC and between-group difference. The changes in CFI were generally not appreciable when comparing constrained model and unconstrained model, which indicated its lack of efficiency to detect between-level noninvariance in factor loadings.

ΔRMSEA. The ΔRMSEA had larger values comparing the level-specific approach with the standard approach. The larger ΔRMSEA values indicated larger model difference and showed to be more sensitive in response to noninvariance in between-level factor loadings. The factors affecting the changes in RMSEA included

ICC and between-group difference. When ICC and between-group difference increased, ΔRMSEA became larger.

In the invariant conditions (100% invariance), ΔRMSEAs were generally very small (e.g., around 0.000). In the noninvariant conditions (0% to 75% invariance), changes in RMSEA were more noticeable with increasing ICC and between-group difference. For example, when ICC and between-group difference were small, ΔRMSEA had small values around 0.000 irrespective of proportion of invariance, number of clusters, and cluster size. ΔRMSEA inflated with the increase in ICC and between-group difference. As shown in Table 4.10, when between-group difference was 0.5, changes in RMSEA ranged from 0.002 to 0.005 and from 0.017 to 0.026 for low ICC and high ICC, respectively, with respect to different degree of noninvariance (100 % to 25%).

In summary, ΔRMSEA appeared to be more sensitive when ICC and between-group difference were large in response to the between-level factor loading noninvariance. The changes of RMSEA were not notably related to the number of cluster and cluster size. In addition, the effect of proportion of invariance on the changes in RMSEA was little. The pattern of changes in RMSEA due to different proportion of invariance was not consistent. In large ICC conditions, when 50% of items were invariant, the changes were the largest, whereas the changes were the smallest when 75% of items were invariant. In general, some changes in RMSEA indicated the difference comparing the metric invariance constrained model to the unconstrained model, which to some extend suggested lack of invariance in between-level factor

loadings. The findings were applicable in large ICC along with large between-group difference conditions.

ΔSRMR. The changes in SRMR were the most striking results in this study. First, the $\Delta SRMR_B$ values in the level-specific approach were found to be identical to $\Delta SRMR_B$ in the standard approach. As discussed previously, $\Delta SRMR_B$ in the standard approach was also a level-specific model fit index for the between-level model. The findings showed the advantage of using level-specific model fit index which listed the model fit of the within-level model and between-level model, respectively. Second, ΔSRMR tended to decrease monotonically as the proportion of invariance decreased. In other words, ΔSRMR exhibited larger values with respect to the increasing degree of noninvariance. Last, ICC and between-group difference affected the changes in SRMR. In addition, ΔSRMR was affected by the number of cluster and cluster size but the effect of cluster size was relatively small.

In the invariant conditions (100% invariance), though ΔSRMR was not zero but was the smallest across all conditions. In the noninvariant conditions, there was an apparent pattern that ΔSRMR tended to be smaller as the degree of invariance increased. For example, as shown in Table 4.14, when between-group difference was 0.5 and ICC was 0.5, changes in SRMR were 0.214, 0.161, 0.148, and 0.078, for 0% invariance, 25% invariance, 50% invariance, and 75% invariance, respectively. This finding meant that SRMR suggested worse fit when the degree of noninvariance increased. To summarize, the changes in SRMR were negatively correlated with all conditions (i.e., ICC, between-group difference, cluster size, number of clusters, and the proportion of invariance). The

property demonstrated that ΔSRMR was a substantial sensitive index to examine invariance in the between-level factor loadings.

ΔTLI. Though not very differently, ΔTLI had larger values in the level-specific approach than in the standard approach. The changes in TLI were affected by ICC and between-group difference. In contrast, the number of cluster, cluster size, and proportion of invariance had smaller effects.

In the invariant conditions, ΔTLIs were generally around 0.000, which indicated metric invariance. In the noninvariant conditions, changes in TLI depended on ICC and between-group difference. When ICC and between-group difference were small, ΔTLI showed little changes comparing the metric invariance constrained model and the unconstrained model. The small changes showed that ΔTLI was not sensitive to detect noninvariant between-level factor loadings. When ICC and between-group difference were large, changes in TLI slightly increased. For example, as is shown in Table 4.13, in conditions of number of cluster of 160, cluster size of 10, and ICC of 0.5, changes in TLI varied from -0.020 to -0.011.

In summary, when ICC and between-group difference were large ΔTLI appeared to be more sensitive in response to the between-level factor loading noninvariance. The effect of different proportion of invariance on the changes in TLI was little. In addition, the pattern of changes in TLI due to different proportion of invariance was not monotonic. For example, in large ICC conditions, when 50% of items were invariant, the changes were the largest, whereas the changes were the smallest when 75% of items were invariant.

TABLE 4.7 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 30 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	14.52	-0.002	0.006	0.040	-0.003	15.33	-0.002	0.005	0.040	-0.002
	25%	14.78	-0.002	0.006	0.037	-0.003	18.34	-0.003	0.005	0.037	-0.003
	50%	12.94	-0.001	0.005	0.035	-0.002	13.34	-0.001	0.004	0.035	-0.002
	75%	11.91	-0.001	0.004	0.032	-0.001	11.91	-0.001	0.003	0.031	-0.001
	100%	11.74	-0.001	0.003	0.028	-0.001	12.39	-0.001	0.003	0.027	-0.001
0.2											
	0%	16.86	-0.002	0.009	0.072	-0.004	18.57	-0.003	0.007	0.072	-0.003
	25%	15.17	-0.002	0.007	0.059	-0.003	15.96	-0.002	0.006	0.059	-0.002
	50%	13.72	-0.001	0.005	0.053	-0.002	15.05	-0.002	0.005	0.054	-0.002
	75%	9.37	0.000	0.001	0.038	0.002	11.67	-0.001	0.002	0.038	-0.001
	100%	6.39	0.000	-0.003	0.024	0.003	7.94	0.000	-0.001	0.024	0.001
0.5											
	0%	28.11	-0.005	0.017	0.173	-0.011	28.19	-0.005	0.013	0.173	-0.006
	25%	35.15	-0.007	0.021	0.130	-0.015	34.74	-0.007	0.016	0.129	-0.009
	50%	42.76	-0.009	0.026	0.125	-0.020	41.39	-0.009	0.019	0.125	-0.011
	75%	28.05	-0.005	0.016	0.069	-0.011	27.50	-0.005	0.012	0.069	-0.006
	100%	8.80	0.000	-0.002	0.023	0.001	9.05	0.000	0.000	0.023	0.000

(continued)

TABLE 4.7 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	13.25	-0.001	0.005	0.032	-0.002	13.47	-0.001	0.004	0.032	-0.002
	25%	12.33	-0.001	0.004	0.031	-0.001	13.38	-0.001	0.004	0.031	-0.002
	50%	10.43	-0.001	0.003	0.030	0.000	11.23	-0.001	0.003	0.030	-0.001
	75%	12.54	-0.001	0.004	0.028	-0.001	13.49	-0.001	0.003	0.028	-0.002
0.2											
	0%	8.83	0.000	0.000	0.036	0.002	10.96	-0.001	0.002	0.035	-0.001
	25%	6.83	0.000	-0.001	0.032	0.003	10.43	-0.001	0.001	0.032	0.000
	50%	8.07	0.000	-0.002	0.029	0.002	7.61	0.000	0.000	0.029	0.001
	75%	5.39	0.001	-0.003	0.026	0.004	7.92	0.000	0.000	0.026	0.001
0.5											
	0%	11.00	-0.001	0.001	0.051	0.000	11.29	-0.001	0.002	0.051	-0.001
	25%	13.75	-0.001	0.004	0.045	-0.002	13.84	-0.001	0.003	0.045	-0.001
	50%	13.79	-0.001	0.004	0.038	-0.002	13.86	-0.001	0.004	0.038	-0.002
	75%	12.01	-0.001	0.002	0.028	-0.001	12.09	-0.001	0.002	0.028	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.8 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 30 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	15.03	-0.001	0.005	0.039	-0.002	15.43	-0.001	0.004	0.039	-0.001
	25%	14.74	-0.001	0.005	0.035	-0.002	14.16	-0.001	0.003	0.035	-0.001
	50%	14.61	-0.001	0.004	0.032	-0.001	14.76	-0.001	0.003	0.032	-0.001
	75%	9.45	0.000	0.002	0.028	0.001	11.60	0.000	0.002	0.028	0.000
	100%	9.25	0.000	0.000	0.022	0.001	10.15	0.000	0.001	0.023	0.000
0.2											
	0%	20.51	-0.002	0.009	0.076	-0.003	20.92	-0.002	0.007	0.076	-0.002
	25%	17.28	-0.001	0.006	0.060	-0.002	17.16	-0.001	0.005	0.060	-0.001
	50%	16.28	-0.001	0.005	0.054	-0.002	15.81	-0.001	0.004	0.054	-0.001
	75%	11.08	0.000	0.001	0.040	0.000	11.70	0.000	0.002	0.040	0.000
	100%	6.51	0.000	-0.003	0.026	0.001	7.41	0.000	-0.001	0.026	0.000
0.5											
	0%	31.01	-0.003	0.014	0.179	-0.006	31.13	-0.003	0.011	0.179	-0.004
	25%	40.18	-0.004	0.018	0.135	-0.009	39.31	-0.004	0.013	0.135	-0.005
	50%	48.95	-0.005	0.022	0.129	-0.012	47.00	-0.005	0.016	0.129	-0.006
	75%	30.49	-0.003	0.014	0.070	-0.006	29.77	-0.003	0.010	0.070	-0.003
	100%	8.96	0.000	-0.001	0.021	0.001	9.02	0.000	0.000	0.021	0.000

(continued)

TABLE 4.8 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	10.69	0.000	0.002	0.027	0.000	12.96	-0.001	0.002	0.027	-0.001
	25%	10.77	0.000	0.002	0.026	0.000	11.98	0.000	0.002	0.026	-0.001
	50%	9.44	0.000	0.001	0.025	0.000	11.07	0.000	0.001	0.025	0.000
	75%	9.62	0.000	0.001	0.024	0.000	10.61	0.000	0.001	0.024	0.000
0.2											
	0%	8.56	0.000	-0.001	0.036	0.001	9.40	0.000	0.000	0.035	0.000
	25%	8.39	0.000	-0.001	0.033	0.001	9.59	0.000	0.000	0.033	0.000
	50%	7.44	0.000	-0.002	0.031	0.001	8.83	0.000	0.000	0.031	0.000
	75%	7.31	0.000	-0.002	0.028	0.001	8.28	0.000	-0.001	0.028	0.000
0.5											
	0%	11.65	-0.001	0.001	0.050	0.000	11.70	-0.001	0.001	0.050	0.000
	25%	14.77	-0.001	0.004	0.044	-0.001	14.71	-0.001	0.003	0.044	-0.001
	50%	14.68	-0.001	0.004	0.037	-0.001	14.59	-0.001	0.003	0.037	-0.001
	75%	12.16	-0.001	0.002	0.026	-0.001	12.14	-0.001	0.002	0.026	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.9 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 50 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	13.70	-0.001	0.005	0.036	-0.002	13.37	-0.001	0.003	0.036	-0.001
	25%	12.57	-0.001	0.004	0.032	-0.001	12.80	-0.001	0.003	0.032	-0.001
	50%	11.48	0.000	0.003	0.030	-0.001	12.40	-0.001	0.003	0.029	-0.001
	75%	10.39	0.000	0.002	0.024	0.000	11.05	0.000	0.001	0.024	-0.001
	100%	6.61	0.000	-0.002	0.019	0.001	6.70	0.000	-0.001	0.019	0.001
0.2											
	0%	20.10	-0.002	0.010	0.077	-0.004	19.20	-0.002	0.007	0.076	-0.002
	25%	17.48	-0.001	0.007	0.059	-0.003	16.43	-0.001	0.005	0.060	-0.002
	50%	16.28	-0.001	0.006	0.054	-0.003	16.43	-0.001	0.005	0.054	-0.002
	75%	11.49	-0.001	0.002	0.038	-0.001	12.26	-0.001	0.002	0.038	-0.001
	100%	6.02	0.000	-0.003	0.023	0.002	6.58	0.000	-0.001	0.023	0.001
0.5											
	0%	37.10	-0.005	0.021	0.188	-0.011	37.15	-0.005	0.015	0.188	-0.006
	25%	50.92	-0.007	0.028	0.140	-0.016	50.09	-0.007	0.020	0.140	-0.009
	50%	62.57	-0.009	0.033	0.132	-0.020	60.88	-0.008	0.024	0.132	-0.011
	75%	39.48	-0.005	0.023	0.071	-0.011	38.76	-0.005	0.016	0.071	-0.006
	100%	8.45	0.000	-0.001	0.017	0.000	8.48	0.000	0.000	0.017	0.000

(continued)

TABLE 4.9 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	9.70	0.000	0.001	0.024	0.000	8.62	0.000	0.001	0.024	0.000
	25%	7.86	0.000	0.000	0.022	0.001	9.31	0.000	0.001	0.022	0.000
	50%	8.24	0.000	0.000	0.021	0.001	8.81	0.000	0.000	0.021	0.000
	75%	7.67	0.000	-0.001	0.019	0.001	8.46	0.000	0.000	0.020	0.000
0.2											
	0%	7.74	0.000	-0.001	0.033	0.001	8.09	0.000	0.000	0.033	0.000
	25%	7.99	0.000	-0.001	0.030	0.001	8.13	0.000	0.000	0.030	0.000
	50%	7.72	0.000	-0.002	0.028	0.001	8.37	0.000	0.000	0.028	0.000
	75%	7.31	0.000	-0.002	0.025	0.001	7.82	0.000	-0.001	0.025	0.000
0.5											
	0%	12.41	-0.001	0.003	0.051	-0.001	12.42	-0.001	0.002	0.051	-0.001
	25%	16.75	-0.001	0.007	0.044	-0.003	16.68	-0.001	0.005	0.044	-0.002
	50%	16.53	-0.001	0.007	0.036	-0.003	16.45	-0.001	0.005	0.036	-0.002
	75%	13.51	-0.001	0.004	0.023	-0.002	13.47	-0.001	0.003	0.023	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.10 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 50 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	14.63	0.000	0.005	0.040	-0.001	16.21	-0.001	0.004	0.040	-0.001
	25%	13.57	0.000	0.003	0.034	-0.001	14.26	0.000	0.003	0.034	-0.001
	50%	11.80	0.000	0.002	0.031	0.000	13.13	0.000	0.002	0.031	-0.001
	75%	7.95	0.000	0.000	0.025	0.000	10.47	0.000	0.001	0.025	0.000
	100%	4.49	0.000	-0.002	0.019	0.001	6.53	0.000	-0.001	0.019	0.000
0.2											
	0%	23.54	-0.001	0.010	0.082	-0.003	23.95	-0.001	0.007	0.082	-0.002
	25%	18.98	-0.001	0.007	0.063	-0.002	19.80	-0.001	0.005	0.063	-0.001
	50%	19.47	-0.001	0.006	0.058	-0.002	19.60	-0.001	0.005	0.058	-0.001
	75%	13.91	0.000	0.003	0.040	-0.001	14.13	0.000	0.002	0.040	-0.001
	100%	7.00	0.000	-0.002	0.023	0.001	7.44	0.000	-0.001	0.023	0.000
0.5											
	0%	44.24	-0.003	0.018	0.192	-0.007	43.38	-0.003	0.013	0.193	-0.004
	25%	58.10	-0.004	0.023	0.144	-0.009	57.09	-0.004	0.016	0.144	-0.005
	50%	71.51	-0.005	0.026	0.135	-0.012	69.43	-0.005	0.019	0.135	-0.006
	75%	42.88	-0.003	0.017	0.071	-0.006	42.11	-0.003	0.013	0.071	-0.003
	100%	8.69	0.000	0.000	0.016	0.000	8.71	0.000	0.000	0.016	0.000

(continued)

TABLE 4.10 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	8.75	0.000	0.000	0.025	0.000	9.28	0.000	0.001	0.025	0.000
	25%	9.34	0.000	0.000	0.023	0.000	9.12	0.000	0.000	0.023	0.000
	50%	6.21	0.000	-0.001	0.021	0.001	8.52	0.000	0.000	0.021	0.000
	75%	6.14	0.000	-0.001	0.020	0.001	7.55	0.000	0.000	0.020	0.000
0.2											
	0%	8.65	0.000	0.000	0.035	0.000	9.34	0.000	0.000	0.035	0.000
	25%	7.76	0.000	0.000	0.032	0.001	9.15	0.000	0.000	0.032	0.000
	50%	8.65	0.000	0.000	0.029	0.000	9.03	0.000	0.000	0.029	0.000
	75%	8.06	0.000	-0.001	0.025	0.000	8.40	0.000	0.000	0.025	0.000
0.5											
	0%	13.24	0.000	0.003	0.050	-0.001	13.21	0.000	0.002	0.050	-0.001
	25%	17.88	-0.001	0.006	0.043	-0.002	17.77	-0.001	0.004	0.043	-0.001
	50%	17.67	-0.001	0.006	0.034	-0.002	17.55	-0.001	0.004	0.034	-0.001
	75%	13.86	0.000	0.003	0.021	-0.001	13.81	0.000	0.002	0.021	-0.001

Note. CN: number of clusters.
CS: cluster size.
B denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.11 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 80 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	14.68	-0.001	0.004	0.037	-0.002	14.15	0.000	0.003	0.036	-0.001
	25%	11.75	0.000	0.003	0.031	-0.001	13.16	0.000	0.002	0.030	-0.001
	50%	10.22	0.000	0.002	0.027	0.000	11.90	0.000	0.002	0.027	0.000
	75%	10.08	0.000	0.001	0.021	0.000	9.85	0.000	0.001	0.021	0.000
	100%	6.55	0.000	-0.002	0.016	0.001	7.24	0.000	-0.001	0.015	0.000
0.2											
	0%	20.84	-0.001	0.010	0.080	-0.003	22.22	-0.001	0.007	0.081	-0.002
	25%	19.68	-0.001	0.008	0.062	-0.003	19.68	-0.001	0.006	0.062	-0.002
	50%	20.66	-0.001	0.008	0.056	-0.003	20.95	-0.001	0.006	0.056	-0.002
	75%	15.22	-0.001	0.005	0.038	-0.002	15.37	-0.001	0.003	0.038	-0.001
	100%	7.25	0.000	-0.001	0.021	0.001	7.49	0.000	-0.001	0.021	0.000
0.5											
	0%	52.28	-0.004	0.024	0.196	-0.011	52.19	-0.004	0.018	0.196	-0.006
	25%	74.68	-0.007	0.032	0.146	-0.016	73.59	-0.007	0.023	0.146	-0.009
	50%	92.61	-0.009	0.037	0.136	-0.020	90.55	-0.008	0.027	0.136	-0.011
	75%	56.10	-0.005	0.026	0.071	-0.011	55.32	-0.005	0.018	0.071	-0.006
	100%	8.60	0.000	0.000	0.013	0.000	8.61	0.000	0.000	0.013	0.000

(continued)

TABLE 4.11 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	8.27	0.000	0.000	0.020	0.000	9.75	0.000	0.001	0.020	0.000
	25%	7.03	0.000	-0.001	0.020	0.001	9.13	0.000	0.000	0.019	0.000
	50%	6.97	0.000	-0.001	0.017	0.001	8.33	0.000	0.000	0.017	0.000
	75%	6.20	0.000	-0.002	0.016	0.001	6.92	0.000	-0.001	0.016	0.000
0.2											
	0%	8.84	0.000	0.000	0.033	0.000	9.04	0.000	0.000	0.033	0.000
	25%	9.40	0.000	0.000	0.032	0.000	9.37	0.000	0.000	0.030	0.000
	50%	9.08	0.000	0.000	0.027	0.000	9.36	0.000	0.000	0.026	0.000
	75%	8.36	0.000	-0.001	0.023	0.000	8.62	0.000	0.000	0.023	0.000
0.5											
	0%	14.70	-0.001	0.004	0.050	-0.001	14.66	-0.001	0.003	0.050	-0.001
	25%	21.27	-0.001	0.009	0.045	-0.003	21.15	-0.001	0.006	0.042	-0.002
	50%	21.19	-0.001	0.009	0.033	-0.003	21.07	-0.001	0.006	0.033	-0.002
	75%	16.30	-0.001	0.006	0.019	-0.002	16.24	-0.001	0.004	0.019	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.12 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 80 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	18.88	-0.001	0.005	0.041	-0.001	18.13	0.000	0.003	0.041	-0.001
	25%	14.72	0.000	0.004	0.034	-0.001	14.34	0.000	0.002	0.034	0.000
	50%	12.65	0.000	0.002	0.030	0.000	12.99	0.000	0.002	0.030	0.000
	75%	8.56	0.000	0.000	0.023	0.000	9.94	0.000	0.001	0.023	0.000
	100%	5.94	0.000	-0.002	0.017	0.000	6.58	0.000	-0.001	0.017	0.000
0.2											
	0%	27.33	-0.001	0.010	0.083	-0.002	26.86	-0.001	0.007	0.083	-0.001
	25%	24.92	-0.001	0.008	0.067	-0.002	25.08	-0.001	0.006	0.067	-0.001
	50%	26.08	-0.001	0.008	0.061	-0.002	25.94	-0.001	0.006	0.061	-0.001
	75%	18.05	-0.001	0.005	0.040	-0.001	18.07	0.000	0.003	0.040	-0.001
	100%	7.87	0.000	-0.001	0.019	0.000	7.97	0.000	0.000	0.019	0.000
0.5											
	0%	57.18	-0.002	0.018	0.187	-0.006	56.68	-0.002	0.013	0.187	-0.003
	25%	86.63	-0.004	0.025	0.152	-0.009	85.36	-0.004	0.018	0.152	-0.005
	50%	105.91	-0.005	0.029	0.141	-0.012	103.40	-0.005	0.021	0.141	-0.006
	75%	61.69	-0.003	0.020	0.073	-0.006	60.77	-0.003	0.014	0.073	-0.003
	100%	8.58	0.000	0.000	0.013	0.000	8.58	0.000	0.000	0.013	0.000

(continued)

TABLE 4.12 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	8.58	0.000	0.000	0.022	0.000	9.58	0.000	0.000	0.022	0.000
	25%	7.31	0.000	-0.001	0.020	0.000	8.16	0.000	0.000	0.020	0.000
	50%	7.15	0.000	-0.001	0.019	0.000	7.46	0.000	0.000	0.019	0.000
	75%	5.87	0.000	-0.001	0.017	0.001	6.68	0.000	-0.001	0.017	0.000
0.2											
	0%	10.53	0.000	0.001	0.035	0.000	10.73	0.000	0.001	0.035	0.000
	25%	11.10	0.000	0.001	0.031	0.000	11.21	0.000	0.001	0.031	0.000
	50%	10.69	0.000	0.001	0.027	0.000	10.76	0.000	0.001	0.027	0.000
	75%	9.36	0.000	0.000	0.022	0.000	9.44	0.000	0.000	0.022	0.000
0.5											
	0%	15.79	0.000	0.004	0.051	-0.001	15.73	0.000	0.003	0.051	0.000
	25%	23.28	-0.001	0.008	0.044	-0.002	23.14	-0.001	0.005	0.044	-0.001
	50%	22.97	-0.001	0.007	0.034	-0.002	22.82	-0.001	0.005	0.034	-0.001
	75%	16.82	0.000	0.004	0.019	-0.001	16.75	0.000	0.003	0.019	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.13 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 160 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	17.48	0.000	0.005	0.041	-0.001	17.58	0.000	0.003	0.041	-0.001
	25%	14.30	0.000	0.004	0.033	-0.001	14.81	0.000	0.002	0.032	-0.001
	50%	14.23	0.000	0.003	0.029	-0.001	14.48	0.000	0.002	0.029	0.000
	75%	11.10	0.000	0.001	0.021	0.000	11.49	0.000	0.001	0.021	0.000
	100%	5.85	0.000	-0.002	0.014	0.000	6.32	0.000	-0.001	0.014	0.000
0.2											
	0%	32.74	-0.001	0.012	0.091	-0.003	32.60	-0.001	0.008	0.091	-0.002
	25%	31.31	-0.001	0.011	0.070	-0.003	30.90	-0.001	0.008	0.070	-0.002
	50%	35.28	-0.001	0.013	0.063	-0.003	35.07	-0.001	0.009	0.063	-0.002
	75%	24.45	-0.001	0.008	0.040	-0.002	24.37	-0.001	0.006	0.040	-0.001
	100%	7.54	0.000	-0.001	0.015	0.000	7.57	0.000	0.000	0.015	0.000
0.5											
	0%	97.17	-0.005	0.028	0.211	-0.011	96.59	-0.005	0.021	0.211	-0.006
	25%	138.45	-0.007	0.036	0.158	-0.016	136.88	-0.007	0.026	0.158	-0.009
	50%	173.58	-0.009	0.041	0.145	-0.020	170.40	-0.008	0.030	0.145	-0.011
	75%	102.58	-0.005	0.029	0.076	-0.011	101.49	-0.005	0.021	0.076	-0.006
	100%	7.84	0.000	0.000	0.009	0.000	7.85	0.000	0.000	0.009	0.000

(continued)

TABLE 4.13 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	7.65	0.000	0.000	0.019	0.000	8.41	0.000	0.000	0.019	0.000
	25%	7.58	0.000	-0.001	0.018	0.000	7.83	0.000	0.000	0.018	0.000
	50%	6.63	0.000	-0.001	0.016	0.000	7.45	0.000	0.000	0.016	0.000
	75%	6.75	0.000	-0.001	0.015	0.000	7.25	0.000	0.000	0.015	0.000
0.2											
	0%	11.53	0.000	0.002	0.035	0.000	11.68	0.000	0.001	0.035	0.000
	25%	12.44	0.000	0.002	0.029	-0.001	12.47	0.000	0.001	0.029	0.000
	50%	12.10	0.000	0.002	0.024	-0.001	12.12	0.000	0.001	0.024	0.000
	75%	10.36	0.000	0.001	0.018	0.000	10.39	0.000	0.001	0.019	0.000
0.5											
	0%	20.46	-0.001	0.007	0.054	-0.002	20.41	-0.001	0.004	0.054	-0.001
	25%	33.25	-0.001	0.012	0.046	-0.003	33.11	-0.001	0.009	0.046	-0.002
	50%	33.36	-0.001	0.012	0.034	-0.003	33.23	-0.001	0.009	0.034	-0.002
	75%	23.42	-0.001	0.008	0.017	-0.002	23.35	-0.001	0.006	0.017	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

TABLE 4.14 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 160 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in factor loadings									
0.1											
	0%	24.59	0.000	0.006	0.044	-0.001	24.14	0.000	0.004	0.044	-0.001
	25%	18.68	0.000	0.004	0.034	-0.001	18.61	0.000	0.003	0.034	0.000
	50%	17.52	0.000	0.003	0.030	-0.001	17.88	0.000	0.002	0.030	0.000
	75%	13.54	0.000	0.002	0.022	0.000	13.66	0.000	0.001	0.022	0.000
	100%	7.21	0.000	-0.001	0.013	0.000	7.33	0.000	0.000	0.013	0.000
0.2											
	0%	45.73	-0.001	0.012	0.095	-0.002	45.81	-0.001	0.008	0.095	-0.001
	25%	41.32	-0.001	0.010	0.074	-0.002	41.17	-0.001	0.007	0.074	-0.001
	50%	43.68	-0.001	0.011	0.065	-0.002	43.43	-0.001	0.008	0.065	-0.001
	75%	28.85	-0.001	0.007	0.040	-0.001	28.75	-0.001	0.005	0.040	-0.001
	100%	7.98	0.000	0.000	0.013	0.000	8.00	0.000	0.000	0.013	0.000
0.5											
	0%	119.60	-0.003	0.023	0.214	-0.007	118.58	-0.003	0.016	0.214	-0.004
	25%	161.19	-0.004	0.028	0.161	-0.009	159.34	-0.004	0.020	0.161	-0.005
	50%	198.48	-0.005	0.031	0.148	-0.011	194.62	-0.005	0.022	0.148	-0.006
	75%	114.91	-0.003	0.022	0.078	-0.006	113.59	-0.003	0.016	0.078	-0.003
	100%	8.24	0.000	0.000	0.009	0.000	8.24	0.000	0.000	0.009	0.000

(continued)

Table 4.14 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in factor loadings									
0.1											
	0%	9.06	0.000	0.000	0.020	0.000	8.68	0.000	0.000	0.020	0.000
	25%	8.85	0.000	0.000	0.018	0.000	9.02	0.000	0.000	0.018	0.000
	50%	8.43	0.000	0.000	0.016	0.000	8.82	0.000	0.000	0.016	0.000
	75%	8.14	0.000	0.000	0.014	0.000	8.29	0.000	0.000	0.014	0.000
0.2											
	0%	14.04	0.000	0.002	0.035	0.000	14.05	0.000	0.001	0.035	0.000
	25%	14.37	0.000	0.002	0.029	0.000	14.37	0.000	0.001	0.029	0.000
	50%	13.59	0.000	0.002	0.023	0.000	13.59	0.000	0.001	0.023	0.000
	75%	11.32	0.000	0.001	0.017	0.000	11.33	0.000	0.001	0.017	0.000
0.5											
	0%	22.44	0.000	0.005	0.054	-0.001	22.38	0.000	0.004	0.054	-0.001
	25%	36.41	-0.001	0.009	0.046	-0.002	36.26	-0.001	0.006	0.046	-0.001
	50%	36.55	-0.001	0.009	0.035	-0.002	36.40	-0.001	0.006	0.035	-0.001
	75%	25.31	0.000	0.006	0.017	-0.001	25.24	0.000	0.004	0.017	-0.001

Note. CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: CFI_B, III: $RMSEA_B$, IV: $SRMR_B$, V: TLI_B.
i: $\Delta\chi^2(8)$, ii: CFI, iii: RMSEA, iv: $SRMR_B$, v: TLI.

Testing Scalar Invariance

Nonconvergence Problems and Improper Solutions. Table 4.15 presented the convergence rates in terms of ICC, number of cluster, and cluster size for testing scalar invariance. As can be seen that only few replications had the convergence problems. The overall convergence rate was 99.91% in testing scalar invariance (intercept invariance). Given small percentages of lack of convergence, nonconverged solutions were excluded before summarizing results. Meanwhile, some replications converged but encountered potential theta problems (negative residual variance estimates). Nevertheless, those replications were included in the following analyses. As discussed previously in metric invariance testing, those replications were not excluded due to two reasons. First, negative residual variance estimates were not significantly different from zero in those replications. Second, after increasing replications from 1,000 to 2,000, model fit indices showed little difference. Therefore, the original results from 1,000 replications were reported only except nonconverged replications were screened out. Similar to the metric invariance test study, replications with extreme chi-square test statistics (e.g., above 2,000) and negative TLI values were also deleted (e.g., one or two out of 1,000 replications when the number of cluster was 30).

Model Fit Indices. The mean GFIs (i.e., CFI, RMSEA, SRMR, and TLI) as well as chi-square statistics with standard deviations were presented in Table 4.16. For X^2 with df of 48 and 88, the critical value is 65.17 and 110.90, respectively. As shown in Table 4.16, mean X^2 was 69.80 (SD = 19.62) and 112.66 (SD = 16.98) for level-specific approach and standard approach, respectively. This finding showed that X^2 overly

TABLE 4.15 Summary of Convergence Rates for Intercept Invariance Testing

		Level-specific approach		Standard approach	
	CN/CS	M1	M2	M3	M4
Small ICC	30/10	98.71%	97.90%	98.59%	97.98%
	30/20	99.88%	99.76%	99.89%	99.81%
	50/10	99.94%	99.89%	99.88%	99.92%
	50/20	100%	100%	100%	100%
	80/10	100%	100%	100%	100%
	80/20	100%	100%	100%	100%
	160/10	100%	100%	100%	100%
	160/20	100%	100%	100%	100%
Average					99.75%
Medium ICC	30/10	99.82%	99.76%	99.83%	99.69%
	30/20	100%	99.99%	100%	100%
	50/10	100%	100%	100%	100%
	50/20	100%	100%	100%	100%
	80/10	100%	100%	100%	100%
	80/20	100%	100%	100%	100%
	160/10	100%	100%	100%	100%
	160/20	100%	100%	100%	100%
Average					99.98%
Large ICC	30/10	100%	100%	99.98%	100%
	30/20	100%	100%	100%	100%
	50/10	100%	100%	100%	100%
	50/20	100%	100%	100%	100%
	80/10	100%	100%	100%	100%
	80/20	100%	100%	100%	100%
	160/10	100%	100%	100%	100%
	160/20	100%	100%	100%	100%
Average					100%

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
M1 & M3: Metric invariance model
M2 & M4: Metric invariance unconstrained model

rejected intercept invariance in 100% invariance conditions. When in noninvariance conditions (0% to 75% invariance), X^2correctly rejected the null hypothesis of intercept invariance. However, X^2 was not stable in response to different degrees of intercept noninvariance. In other words, $X^2{}_B$ did not decrease as the degree of invariance increased regardless of the level-specific approach or the standard approach.

In the level specific approach, the average CFI_B ranged from 0.995 to 0.999 with small SDs across different proportions of invariance (0% to 100% invariance), all suggesting good model fit. It was concluded that CFI_B was not able to detect intercept noninvariance. Similarly, TLI_B with a range of 0.993 to 0.998 indicated good fit. This finding showed that TLI_B neither was sensitive to lack of intercept noninvariance. $RMSEA_B$ was not a good performing index either with all values small than the suggested cutoff of 0.08 though noninvariance was present. In 100% invariance conditions, $RMSEA_B$ was smaller than the cutoff, correctly indicating scalar invariance. In the standard approach, both CFI and TLI suggested good model fit irrespective of the proportion of invariance ranging from 0% to 100%. This results indicated that neither CFI nor TLI was sensitive to lack of intercept noninvariance. In invariant conditions, RMSEA indicated scalar invariance, whereas in noninvariant conditions, RMSEA failed to detect noninvariant intercepts.

$SRMR_B$ had identical results comparing level-specific approach and standard approach. $SRMR_B$ was able to detect noninvariance in between-level intercepts, as shown in Table 4.16 with all values larger than the suggested cutoff of 0.08. In conditions of 100% invariance, SRMRB was around the cutoff (i.e., 0.086).

Next, a five-way analysis of variance (ANOVA) was conducted to examine the factors associated with changes in each fit index separately. The results were presented in Table 4.17 and Table 4.18 for the level-specific approach and the standard approach, respectively. Specifically, the factors were different proportions of invariance (INV), ICC, number of clusters (CN), cluster sizes (CS), and between-group difference (DIF). Eta-squared effects sizes (η^2) were used to quantify the effects.

For the level-specific approach, DIF was the most influential factor that affected the changes in all GFIs and X^2. η^2 were 34.68%, 39.97%, 43.06%, 22.05%, and 40.41% for ΔX^2, ΔCFI, ΔRMSEA, ΔSRMR, and ΔTLI, respectively. Number of clusters had the second largest effect on the change of X^2 (η^2 = 14.80%). Nevertheless, for other indices (i.e., ΔCFI, ΔRMSEA, ΔSRMR, and ΔTLI), ICC was the second influential factor. Similarly, for the standard approach DIF had the largest effect on the changes in the five indices investigated (i.e., ΔX^2, ΔCFI, ΔRMSEA, ΔSRMR, and ΔTLI). The factor with the second largest effect was ICC. To summarize, between-group difference and ICC accounted for a large proportion of variance in the changes of fit indices investigated in this study.

Fit Indices Sensitivity in Scalar Invariance Tests. The changes in five fit indices investigated in this study (i.e., ΔX^2,ΔCFI, ΔRMSEA, ΔSRMR, and ΔTLI) were summarized in Table 4.18 to Table 4.25 in terms of different number of clusters and cluster sizes. As can be seen from the results, ΔGFIs and ΔX^2 in response to lack of intercept invariance showed similar pattern to results of examining factor loading invariance in between-level models, which were detailed as follows.

TABLE 4.16 Means and SDs (in parenthesis) of Model Fit Indices for Scalar Invariance Models

Invariance (%)	Level-specific Approach [a]					"Standard" Approach				
	$\chi^2(48)_B$	CFI_B	$RMSEA_B$	$SRMR_B$	TLI_B	$\chi^2(88)$	CFI	RMSEA	$SRMR_B$	TLI
0%	94.31	0.997	0.023	0.106	0.993	140.07	0.996	0.017	0.106	0.996
	(34.40)	(0.003)	(0.013)	(0.054)	(0.007)	(34.93)	(0.004)	(0.010)	(0.054)	(0.004)
25%	115.46	0.995	0.029	0.100	0.990	160.97	0.995	0.022	0.100	0.995
	(49.61)	(0.004)	(0.014)	(0.049)	(0.009)	(50.06)	(0.004)	(0.011)	(0.049)	(0.005)
50%	117.44	0.995	0.030	0.093	0.990	162.81	0.995	0.022	0.093	0.994
	(51.98)	(0.004)	(0.015)	(0.045)	(0.009)	(52.48)	(0.005)	(0.011)	(0.045)	(0.005)
75%	96.75	0.996	0.024	0.088	0.993	142.48	0.996	0.018	0.088	0.996
	(35.65)	(0.004)	(0.014)	(0.043)	(0.007)	(35.73)	(0.004)	(0.010)	(0.043)	(0.004)
100%	69.80	0.999	0.012	0.086	0.998	112.66	0.998	0.010	0.086	0.998
	(19.62)	(0.002)	(0.012)	(0.043)	(0.005)	(16.98)	(0.003)	(0.010)	(0.043)	(0.003)

Note. [a] In the level-specific approach, the within-level model is saturated.

$_B$ denotes between-level.

CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean square residual; TLI = the Tucker-Lewis index.

TABLE 4.17 Eta-Squared Effect Size Estimates for Study Variables on the Change in Fit Indices (Level-specific Indices)

Variables	Level-specific Indices				
	$\Delta\chi^2(8)_B$	ΔCFI_B	$\Delta RMSEA_B$	$\Delta SRMR_B$	ΔTLI_B
INV	0.46%	0.78%	0.50%	5.48%	0.77%
ICC	4.63%	5.44%	4.81%	12.96%	5.46%
CN	14.80%	0.06%	2.65%	0.28%	0.01%
CS	0.04%	4.33%	1.30%	0.00%	4.34%
DIF	34.68%	39.97%	43.06%	22.05%	40.41%
INV * ICC	0.75%	0.72%	1.04%	5.01%	0.75%
INV * CN	0.19%	0.00%	0.02%	0.23%	0.00%
INV * CS	0.01%	0.08%	0.00%	0.01%	0.08%
INV * DIF	0.32%	0.63%	0.16%	8.13%	0.59%
ICC * CN	1.60%	0.01%	0.06%	0.39%	0.01%
ICC * CS	0.00%	0.64%	0.16%	0.02%	0.65%
ICC * DIF	2.16%	2.70%	1.14%	9.95%	2.52%
CN * CS	0.03%	0.02%	0.14%	0.01%	0.00%
CN * DIF	12.27%	0.00%	0.60%	0.49%	0.01%
CS * DIF	0.03%	4.01%	1.01%	0.00%	4.10%
INV * ICC * CN	0.28%	0.00%	0.02%	0.20%	0.00%
INV * ICC * CS	0.01%	0.04%	0.01%	0.02%	0.04%
INV * ICC * DIF	0.35%	0.35%	0.30%	3.98%	0.33%
INV * CN * CS	0.00%	0.00%	0.00%	0.01%	0.00%
INV * CN * DIF	0.11%	0.00%	0.00%	0.25%	0.00%
INV * CS * DIF	0.00%	0.07%	0.00%	0.00%	0.06%
ICC * CN * CS	0.00%	0.00%	0.01%	0.00%	0.00%
ICC * CN * DIF	0.76%	0.00%	0.01%	0.29%	0.01%
ICC * CS * DIF	0.00%	0.32%	0.04%	0.01%	0.29%

(continued)

TABLE 4.17 (continued)

Variables	Level-specific Indices				
	$\Delta\chi^2(8)_B$	ΔCFI_B	$\Delta RMSEA_B$	$\Delta SRMR_B$	ΔTLI_B
CN * CS * DIF	0.01%	0.00%	0.03%	0.00%	0.00%
INV * ICC * CN * CS	0.00%	0.00%	0.00%	0.00%	0.00%
INV * ICC * CN * DIF	0.13%	0.00%	0.00%	0.12%	0.00%
INV * ICC * CS * DIF	0.00%	0.02%	0.00%	0.01%	0.02%
INV * CN * CS * DIF	0.00%	0.00%	0.00%	0.01%	0.00%
ICC * CN * CS * DIF	0.00%	0.00%	0.01%	0.00%	0.01%
INV * ICC * CN * CS * DIF	0.00%	0.00%	0.00%	0.00%	0.00%

Note: INV = Invariance proportion. ICC = Intra-class correlation. CN = Number of clusters. CS = Cluster size. DIF = Between-group difference.
$_B$ denotes between-level model.
CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean square residual; TLI = the Tucker-Lewis index.

TABLE 4.18 Eta-Squared Effect Size Estimates for Study Variables on the Change in Fit Indices (Standard Indices)

Variables	Standard Indices				
	$\Delta\chi^2(8)$	ΔCFI	ΔRMSEA	$\Delta SRMR_B$	ΔTLI
INV	0.48%	0.78%	0.51%	21.54%	0.11%
ICC	4.34%	5.18%	4.43%	50.67%	0.77%
CN	16.72%	0.11%	3.28%	1.12%	0.01%
CS	0.22%	3.85%	0.72%	0.02%	0.57%
DIF	35.53%	40.71%	44.58%	87.02%	6.13%
INV * ICC	0.75%	0.72%	1.05%	19.60%	0.11%
INV * CN	0.20%	0.00%	0.02%	0.93%	0.00%
INV * CS	0.01%	0.07%	0.00%	0.04%	0.01%
INV * DIF	0.33%	0.64%	0.17%	32.07%	0.09%
ICC * CN	1.65%	0.04%	0.14%	1.53%	0.01%
ICC * CS	0.02%	0.56%	0.12%	0.07%	0.08%
ICC * DIF	2.36%	2.88%	1.35%	39.21%	0.40%
CN * CS	0.16%	0.04%	0.06%	0.02%	0.00%
CN * DIF	12.26%	0.01%	0.73%	1.96%	0.00%
CS * DIF	0.01%	4.26%	1.23%	0.00%	0.65%
INV * ICC * CN	0.28%	0.00%	0.02%	0.77%	0.00%
INV * ICC * CS	0.01%	0.04%	0.01%	0.08%	0.01%
INV * ICC * DIF	0.35%	0.35%	0.30%	15.66%	0.05%
INV * CN * CS	0.01%	0.00%	0.00%	0.05%	0.00%
INV * CN * DIF	0.11%	0.00%	0.00%	0.98%	0.00%
INV * CS * DIF	0.00%	0.07%	0.00%	0.01%	0.01%
ICC * CN * CS	0.05%	0.03%	0.10%	0.02%	0.01%
ICC * CN * DIF	0.82%	0.04%	0.12%	1.14%	0.01%
ICC * CS * DIF	0.02%	0.38%	0.10%	0.03%	0.05%

(continued)

TABLE 4.18 (continued)

Variables	Standard Indices				
	$\Delta\chi^2(8)$	ΔCFI	ΔRMSEA	$\Delta SRMR_B$	ΔTLI
CN * CS * DIF	0.03%	0.01%	0.08%	0.02%	0.00%
INV * ICC * CN * CS	0.00%	0.00%	0.00%	0.01%	0.00%
INV * ICC * CN * DIF	0.13%	0.00%	0.00%	0.49%	0.00%
INV * ICC * CS * DIF	0.00%	0.02%	0.00%	0.03%	0.00%
INV * CN * CS * DIF	0.00%	0.00%	0.00%	0.03%	0.00%
ICC * CN * CS * DIF	0.06%	0.04%	0.11%	0.02%	0.01%
INV * ICC * CN * CS * DIF	0.00%	0.00%	0.00%	0.01%	0.00%

Note: INV = Invariance proportion. ICC = Intra-class correlation. CN = Number of clusters. CS = Cluster size. DIF = Between-group difference.
$_B$ denotes between-level model.
CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean square residual; TLI = the Tucker-Lewis index.

ΔX². The ΔX^2 values in the level-specific approach and in the standard approach were discussed together because the results were similar. As factors identified in ANOVA, the changes in X^2were related to between-group difference, ICC and the number of clusters. ΔX^2 tended to be larger when between-group difference increased. ΔX^2 was greater if ICC was smaller, which was contrary to the results when examining factor loading invariance. When the number of clusters increased, the changes in X^2 were larger.

In the invariant conditions (100% invariance), all ΔX^2values were below the critical value (χ^2_{critical} (0.05, 8) = 15.50), indicating scalar invariance. In the noninvariant conditions (0% to 75% invariance), when between-group difference was large, $\Delta X^2$was able to detect between-level noninvariant intercepts. When between-group difference was small, large number of clusters was required for ΔX^2 to respond to lack of intercept invariance. It was noted that the changes in ΔX^2were not consistent. For example, when 50% of items were invariant (i.e., 50% invariance), $\Delta\chi^2$ was the largest, whereas invariant items increased to 75% (i.e., 75% invariance), $\Delta\chi^2$ was the smallest.

ΔCFI. Similarly, between-group difference and ICC affected the changes in CFI. However, the effect of number of clusters was negligible while the cluster sizes had some effects. There was no much difference comparing the level-specific approach with the standard approach. When between-group difference was small, changes in CFI approximated zero. With increasing between-group difference and the number of cluster, ΔCFI was larger. Changes in CFI were smaller given the increase in ICC.

In the invariant conditions, ΔCFI was nearly all zero, which indicated intercept invariance. In the noninvariant conditions, when between-group difference was large, ΔCFI ranged from -0.010 to -0.002. When between-group difference was small, ΔCFI had a small range (e.g., -0.002 to -0.001). Nonmonotonic changes in CFI were observed with respect to different proportions of invariance. For example, when 50% of items were invariant (i.e., 50% invariance), ΔCFI was the largest, whereas invariant items increased to 75% (i.e., 75% invariance), ΔCFI was the smallest.

ΔRMSEA. The changes in RMSEA were relatively larger in the level-specific approach than in the standard approach (i.e., $\Delta RMSEA_B > \Delta RMSEA$). This finding showed that $\Delta RMSEA_B$ in level-specific approach was more sensitive to noninvariant intercepts than general ΔRMSEA for the whole model. The changes in RMSEA were largely affected by between-group difference and ICC. In addition, number of cluster and cluster size had some effects. Larger ΔRMSEA values in large between-group difference conditions indicated that ΔRMSEA was more sensitive to detect noninvariant intercepts compared to ΔRMSEA in small between-group difference conditions. While ICC increased, ΔRMSEA tinily deceased. ΔRMSEA in conditions of larger number of clusters and cluster sizes had mild increase than in small number of clusters and cluster sizes.

In the invariant conditions, ΔRMSEA was around zero, which indicated intercept invariance. In the noninvariant conditions, ΔRMSEA was larger when between-group difference was larger. The changes in RMSEA were not monotonic due to different proportions of invariance. For example, when 50% of items were invariant (i.e., 50%

invariance), ΔRMSEA was the largest, whereas invariant items raised to 75% (i.e., 75% invariance), ΔRMSEA was the smallest. On the other hand, the changes in RMSEA showed a similar pattern due to different proportions of invariance when ICC changed (small ICC to large ICC).

ΔSRMR. The changes in $SRMR_B$ showed a very regular pattern. First, $\Delta SRMR_B$ in the level-specific approach was identical to $\Delta SRMR_B$ in the standard approach. As explained previously, $\Delta SRMR_B$ in the standard approach was also a fit index for the between-level model only. Second, $\Delta SRMR_B$ exhibited an obviously monotonic change due to different degrees of invariance when between-group difference was large. Specifically, when the proportions of invariance increased, $\Delta SRMR_B$ tended to be smaller. In other words, $\Delta SRMR_B$ reflected the influence of the degree of noninvariance in a consistent way. Last, ICC influenced the changes in $SRMR_B$ along with the effects from the number of cluster and cluster size.

In the invariant conditions, $\Delta SRMR_B$ was zero, indicating intercept invariance. In the noninvariant conditions, noticeably $\Delta SRMR_B$ tended to be smaller as the degree of invariance increased, though this pattern was less apparent in small between-group difference conditions. For example, as shown in Table 4.25, when between-group difference was 0.5 and ICC was small, changes in $SRMR_B$ were 0.075, 0.051, 0.030, and 0.004, respectively, when the proportions of invariance changed from 0% to 75%. This finding demonstrated that $SRMR_B$ was sensitive to lack of invariance in between-level intercepts. However, the changes was not that obvious when the between-group

difference was small. To summarize, in examining between-level intercept invariance, $\Delta SRMR_B$ was noninvariance susceptible when between-group difference was large.

ΔTLI. In general, ΔTLI had larger values in the level-specific approach than in the standard approach. The changes in TLI were affected largely by between-group difference and ICC levels.

In the invariant conditions, ΔTLI was generally around 0.000. In the noninvariant conditions, changes in TLI appeared to depend on between-group difference. When between-group difference was large, ΔTLI showed more changes relative to small between-group conditions. Again, proportion of invariance affected ΔTLI neither consistently nor substantially. For example, in 50% invariance conditions, the changes were the largest, whereas in 75% invariance conditions, the changes were the smallest.

TABLE 4.19 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 30 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	33.60	-0.007	0.019	0.045	-0.013	33.88	-0.007	0.014	0.045	-0.007
	25%	43.71	-0.010	0.025	0.029	-0.018	43.08	-0.010	0.018	0.029	-0.010
	50%	44.23	-0.010	0.025	0.016	-0.019	42.73	-0.009	0.018	0.016	-0.010
	75%	31.40	-0.006	0.017	0.005	-0.012	30.22	-0.006	0.012	0.005	-0.006
	100%	6.42	0.000	-0.002	0.000	0.003	8.14	0.000	-0.001	0.000	0.000
0.2											
	0%	25.87	-0.005	0.013	0.032	-0.008	26.01	-0.005	0.010	0.032	-0.005
	25%	40.24	-0.009	0.023	0.022	-0.016	39.65	-0.008	0.017	0.022	-0.009
	50%	42.66	-0.009	0.024	0.012	-0.018	42.61	-0.009	0.018	0.012	-0.010
	75%	30.27	-0.006	0.016	0.004	-0.011	29.92	-0.006	0.012	0.004	-0.006
	100%	8.53	0.000	-0.002	0.000	0.001	8.89	0.000	0.000	0.000	0.000
0.5											
	0%	17.72	-0.002	0.007	0.007	-0.004	17.69	-0.002	0.005	0.007	-0.002
	25%	37.67	-0.008	0.022	0.006	-0.014	37.24	-0.007	0.016	0.006	-0.008
	50%	41.62	-0.009	0.024	0.004	-0.016	41.06	-0.008	0.017	0.004	-0.009
	75%	29.48	-0.005	0.016	0.002	-0.010	29.15	-0.005	0.012	0.002	-0.006
	100%	8.72	0.000	-0.002	0.001	0.001	8.81	0.000	0.000	0.001	0.000

(continued)

TABLE 4.19 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	12.87	-0.001	0.003	0.001	-0.001	13.87	-0.002	0.003	0.001	-0.001
	25%	14.58	-0.002	0.005	0.001	-0.002	15.10	-0.002	0.004	0.001	-0.002
	50%	12.89	-0.001	0.004	0.001	-0.001	14.40	-0.002	0.004	0.001	-0.002
	75%	10.07	-0.001	0.001	0.000	0.001	12.15	-0.001	0.002	0.000	-0.001
0.2											
	0%	11.44	-0.001	0.001	0.003	0.000	11.73	-0.001	0.002	0.003	-0.001
	25%	14.23	-0.002	0.004	0.002	-0.002	14.43	-0.002	0.003	0.002	-0.002
	50%	14.38	-0.002	0.004	0.002	-0.002	14.63	-0.002	0.003	0.002	-0.002
	75%	12.09	-0.001	0.002	0.001	-0.001	12.43	-0.001	0.002	0.001	-0.001
0.5											
	0%	10.27	-0.001	0.000	0.001	0.000	10.34	-0.001	0.001	0.001	0.000
	25%	13.77	-0.001	0.004	0.001	-0.002	13.78	-0.001	0.003	0.001	-0.001
	50%	14.26	-0.002	0.004	0.001	-0.002	14.26	-0.002	0.003	0.001	-0.001
	75%	12.21	-0.001	0.002	0.001	-0.001	12.24	-0.001	0.002	0.001	-0.001

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.20 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 30 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	36.61	-0.004	0.015	0.052	-0.007	36.49	-0.004	0.011	0.052	-0.004
	25%	45.28	-0.005	0.019	0.034	-0.009	44.79	-0.005	0.014	0.034	-0.005
	50%	44.93	-0.005	0.018	0.017	-0.009	43.81	-0.005	0.013	0.017	-0.005
	75%	31.98	-0.003	0.013	0.004	-0.006	31.83	-0.003	0.009	0.004	-0.003
	100%	8.45	0.000	-0.001	0.000	0.001	8.23	0.000	-0.001	0.001	0.000
0.2											
	0%	28.22	-0.003	0.011	0.033	-0.005	28.12	-0.003	0.008	0.033	-0.003
	25%	42.20	-0.005	0.017	0.023	-0.009	41.76	-0.004	0.013	0.023	-0.005
	50%	44.24	-0.005	0.018	0.010	-0.009	43.49	-0.005	0.013	0.010	-0.005
	75%	30.98	-0.003	0.012	0.003	-0.006	30.51	-0.003	0.009	0.003	-0.003
	100%	8.04	0.000	-0.001	0.001	0.001	8.22	0.000	-0.001	0.001	0.000
0.5											
	0%	17.70	-0.001	0.005	0.007	-0.002	17.65	-0.001	0.004	0.007	-0.001
	25%	38.93	-0.004	0.016	0.006	-0.007	38.45	-0.004	0.012	0.006	-0.004
	50%	42.84	-0.005	0.018	0.004	-0.009	42.26	-0.004	0.013	0.004	-0.005
	75%	30.02	-0.003	0.012	0.002	-0.005	29.69	-0.003	0.009	0.002	-0.003
	100%	8.25	0.000	-0.001	0.000	0.001	8.33	0.000	0.000	0.000	0.000

(continued)

TABLE 4.20 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	13.66	-0.001	0.003	0.003	-0.001	13.87	-0.001	0.002	0.003	-0.001
	25%	16.45	-0.001	0.004	0.002	-0.002	15.63	-0.001	0.003	0.002	-0.001
	50%	14.76	-0.001	0.003	0.001	-0.001	14.80	-0.001	0.003	0.001	-0.001
	75%	13.15	-0.001	0.002	0.001	-0.001	12.76	-0.001	0.002	0.001	-0.001
0.2											
	0%	11.72	-0.001	0.001	0.003	0.000	11.84	-0.001	0.001	0.003	0.000
	25%	14.62	-0.001	0.003	0.002	-0.001	14.66	-0.001	0.003	0.002	-0.001
	50%	14.48	-0.001	0.003	0.001	-0.001	14.51	-0.001	0.003	0.001	-0.001
	75%	11.96	-0.001	0.001	0.001	0.000	12.03	-0.001	0.001	0.001	0.000
0.5											
	0%	9.88	0.000	0.000	0.001	0.000	9.93	0.000	0.000	0.001	0.000
	25%	13.69	-0.001	0.003	0.001	-0.001	13.69	-0.001	0.002	0.001	-0.001
	50%	14.15	-0.001	0.003	0.001	-0.001	14.14	-0.001	0.003	0.001	-0.001
	75%	11.95	-0.001	0.002	0.001	0.000	11.97	-0.001	0.001	0.001	0.000

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.21 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 50 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	48.11	-0.007	0.024	0.055	-0.013	47.68	-0.007	0.018	0.055	-0.007
	25%	63.65	-0.009	0.031	0.037	-0.019	62.83	-0.009	0.022	0.037	-0.010
	50%	63.96	-0.009	0.031	0.020	-0.019	62.26	-0.009	0.022	0.020	-0.010
	75%	44.47	-0.006	0.022	0.004	-0.012	43.15	-0.006	0.016	0.004	-0.006
	100%	8.15	0.000	-0.001	0.000	0.001	8.48	0.000	0.000	0.000	0.000
0.2											
	0%	37.58	-0.005	0.020	0.036	-0.010	37.34	-0.005	0.014	0.036	-0.005
	25%	59.76	-0.009	0.030	0.025	-0.017	59.16	-0.008	0.021	0.025	-0.009
	50%	62.39	-0.009	0.031	0.011	-0.018	61.64	-0.009	0.022	0.011	-0.010
	75%	42.18	-0.006	0.022	0.002	-0.011	41.61	-0.005	0.016	0.002	-0.006
	100%	8.31	0.000	-0.001	0.000	0.000	8.36	0.000	0.000	0.000	0.000
0.5											
	0%	22.75	-0.002	0.011	0.007	-0.004	22.66	-0.002	0.008	0.007	-0.002
	25%	54.74	-0.007	0.028	0.006	-0.014	54.27	-0.007	0.020	0.006	-0.008
	50%	60.01	-0.008	0.030	0.003	-0.016	59.45	-0.008	0.021	0.003	-0.009
	75%	40.68	-0.005	0.021	0.002	-0.010	40.36	-0.005	0.015	0.002	-0.005
	100%	8.39	0.000	-0.001	0.000	0.000	8.42	0.000	0.000	0.000	0.000

(continued)

TABLE 4.21 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	16.31	-0.001	0.006	0.002	-0.002	16.19	-0.001	0.004	0.002	-0.001
	25%	19.32	-0.002	0.008	0.002	-0.003	19.04	-0.002	0.006	0.002	-0.002
	50%	18.55	-0.002	0.007	0.001	-0.003	18.32	-0.002	0.005	0.001	-0.002
	75%	11.23	-0.001	0.001	0.000	-0.001	11.35	-0.001	0.001	0.000	-0.001
0.2											
	0%	13.56	-0.001	0.004	0.003	-0.001	13.55	-0.001	0.003	0.003	-0.001
	25%	17.91	-0.002	0.007	0.002	-0.003	17.83	-0.001	0.005	0.002	-0.002
	50%	17.54	-0.001	0.007	0.001	-0.003	17.46	-0.001	0.005	0.001	-0.002
	75%	11.08	-0.001	0.001	0.000	-0.001	11.10	-0.001	0.001	0.000	0.000
0.5											
	0%	10.81	0.000	0.001	0.001	-0.001	10.81	0.000	0.001	0.001	0.000
	25%	16.46	-0.001	0.006	0.001	-0.002	16.42	-0.001	0.004	0.001	-0.001
	50%	16.87	-0.001	0.006	0.001	-0.002	16.82	-0.001	0.005	0.001	-0.001
	75%	11.02	0.000	0.001	0.000	-0.001	11.03	0.000	0.001	0.000	0.000

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.22 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 50 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	54.59	-0.004	0.020	0.060	-0.008	54.23	-0.004	0.014	0.060	-0.004
	25%	67.27	-0.005	0.023	0.040	-0.010	66.64	-0.005	0.017	0.040	-0.005
	50%	66.68	-0.005	0.023	0.021	-0.010	65.43	-0.005	0.016	0.021	-0.005
	75%	45.36	-0.003	0.017	0.004	-0.006	44.31	-0.003	0.012	0.004	-0.003
	100%	8.00	0.000	-0.001	0.000	0.000	8.03	0.000	0.000	0.000	0.000
0.2											
	0%	41.00	-0.003	0.015	0.037	-0.005	40.77	-0.003	0.011	0.037	-0.003
	25%	62.42	-0.004	0.022	0.025	-0.009	61.92	-0.004	0.016	0.025	-0.005
	50%	65.19	-0.005	0.023	0.011	-0.009	64.42	-0.005	0.016	0.011	-0.005
	75%	43.55	-0.003	0.016	0.002	-0.006	43.07	-0.003	0.012	0.002	-0.003
	100%	8.22	0.000	-0.001	0.000	0.000	8.26	0.000	0.000	0.000	0.000
0.5											
	0%	23.40	-0.001	0.008	0.007	-0.002	23.32	-0.001	0.006	0.007	-0.001
	25%	56.52	-0.004	0.020	0.006	-0.008	56.06	-0.004	0.015	0.006	-0.004
	50%	62.76	-0.004	0.022	0.003	-0.009	62.20	-0.004	0.016	0.003	-0.005
	75%	42.00	-0.003	0.016	0.002	-0.005	41.66	-0.003	0.011	0.002	-0.003
	100%	8.25	0.000	-0.001	0.000	0.000	8.28	0.000	0.000	0.000	0.000

(continued)

TABLE 4.22 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	17.16	-0.001	0.005	0.004	-0.001	17.09	-0.001	0.003	0.004	-0.001
	25%	20.14	-0.001	0.006	0.003	-0.002	19.91	-0.001	0.005	0.003	-0.001
	50%	19.01	-0.001	0.006	0.001	-0.002	18.83	-0.001	0.004	0.001	-0.001
	75%	14.32	-0.001	0.003	0.001	-0.001	14.25	0.000	0.002	0.001	-0.001
0.2											
	0%	14.29	0.000	0.003	0.003	-0.001	14.27	0.000	0.002	0.003	-0.001
	25%	18.49	-0.001	0.006	0.002	-0.002	18.41	-0.001	0.004	0.002	-0.001
	50%	18.30	-0.001	0.006	0.001	-0.002	18.21	-0.001	0.004	0.001	-0.001
	75%	14.12	0.000	0.003	0.001	-0.001	14.09	0.000	0.002	0.001	-0.001
0.5											
	0%	10.88	0.000	0.001	0.001	0.000	10.89	0.000	0.001	0.001	0.000
	25%	16.73	-0.001	0.005	0.001	-0.001	16.68	-0.001	0.003	0.001	-0.001
	50%	17.46	-0.001	0.005	0.001	-0.001	17.41	-0.001	0.004	0.001	-0.001
	75%	13.81	0.000	0.003	0.000	-0.001	13.79	0.000	0.002	0.000	0.000

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.23 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 80 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	70.60	-0.007	0.028	0.065	-0.014	70.32	-0.006	0.020	0.065	-0.007
	25%	94.26	-0.009	0.034	0.044	-0.019	93.57	-0.009	0.025	0.044	-0.010
	50%	94.24	-0.009	0.034	0.025	-0.019	92.56	-0.009	0.025	0.025	-0.010
	75%	64.75	-0.006	0.026	0.005	-0.012	62.34	-0.006	0.018	0.005	-0.006
	100%	8.10	0.000	-0.001	0.000	0.000	7.96	0.000	0.000	0.000	0.000
0.2											
	0%	54.60	-0.005	0.023	0.041	-0.010	54.34	-0.005	0.017	0.040	-0.005
	25%	88.93	-0.008	0.033	0.028	-0.017	88.34	-0.008	0.024	0.028	-0.009
	50%	92.92	-0.009	0.034	0.012	-0.018	92.01	-0.009	0.025	0.012	-0.010
	75%	61.23	-0.006	0.025	0.002	-0.011	60.65	-0.005	0.018	0.002	-0.006
	100%	8.12	0.000	-0.001	0.000	0.000	8.15	0.000	0.000	0.000	0.000
0.5											
	0%	30.80	-0.002	0.014	0.008	-0.004	30.70	-0.002	0.009	0.008	-0.002
	25%	80.81	-0.007	0.031	0.007	-0.015	80.35	-0.007	0.022	0.007	-0.008
	50%	89.73	-0.008	0.034	0.003	-0.016	89.17	-0.008	0.024	0.003	-0.009
	75%	59.08	-0.005	0.025	0.002	-0.010	58.72	-0.005	0.018	0.002	-0.006
	100%	8.15	0.000	-0.001	0.000	0.000	8.17	0.000	0.000	0.000	0.000

(continued)

TABLE 4.23 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	19.87	-0.001	0.008	0.004	-0.002	19.83	-0.001	0.005	0.004	-0.001
	25%	25.11	-0.002	0.011	0.003	-0.004	25.03	-0.002	0.007	0.003	-0.002
	50%	24.43	-0.002	0.010	0.001	-0.003	24.38	-0.002	0.007	0.001	-0.002
	75%	17.41	-0.001	0.006	0.000	-0.002	17.23	-0.001	0.004	0.000	-0.001
0.2											
	0%	16.56	-0.001	0.005	0.003	-0.002	16.52	-0.001	0.004	0.003	-0.001
	25%	23.31	-0.001	0.010	0.002	-0.003	23.22	-0.001	0.007	0.002	-0.002
	50%	23.00	-0.001	0.010	0.001	-0.003	22.91	-0.001	0.006	0.001	-0.002
	75%	17.01	-0.001	0.006	0.000	-0.002	16.95	-0.001	0.004	0.000	-0.001
0.5											
	0%	11.99	0.000	0.002	0.001	-0.001	11.99	0.000	0.002	0.001	0.000
	25%	20.91	-0.001	0.008	0.001	-0.002	20.86	-0.001	0.006	0.001	-0.001
	50%	21.86	-0.001	0.009	0.000	-0.003	21.80	-0.001	0.006	0.000	-0.001
	75%	16.59	-0.001	0.005	0.000	-0.002	16.55	-0.001	0.004	0.000	-0.001

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.24 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 80 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	80.84	-0.004	0.022	0.066	-0.008	80.47	-0.004	0.016	0.066	-0.004
	25%	100.04	-0.005	0.025	0.044	-0.010	99.45	-0.005	0.018	0.044	-0.005
	50%	97.88	-0.005	0.025	0.023	-0.009	96.74	-0.005	0.018	0.023	-0.005
	75%	66.34	-0.003	0.019	0.003	-0.006	65.27	-0.003	0.013	0.003	-0.003
	100%	8.02	0.000	0.000	0.000	0.000	8.05	0.000	0.000	0.000	0.000
0.2											
	0%	58.65	-0.003	0.017	0.040	-0.005	58.42	-0.003	0.012	0.040	-0.003
	25%	91.94	-0.004	0.024	0.027	-0.009	91.47	-0.004	0.017	0.027	-0.005
	50%	95.97	-0.005	0.025	0.011	-0.009	95.22	-0.004	0.018	0.011	-0.005
	75%	63.61	-0.003	0.018	0.002	-0.006	63.11	-0.003	0.013	0.002	-0.003
	100%	8.06	0.000	0.000	0.000	0.000	8.08	0.000	0.000	0.000	0.000
0.5											
	0%	31.14	-0.001	0.010	0.007	-0.002	31.06	-0.001	0.007	0.007	-0.001
	25%	82.79	-0.004	0.022	0.007	-0.007	82.36	-0.004	0.016	0.007	-0.004
	50%	92.64	-0.004	0.024	0.003	-0.008	92.09	-0.004	0.017	0.003	-0.005
	75%	61.58	-0.003	0.018	0.002	-0.005	61.24	-0.003	0.013	0.002	-0.003
	100%	8.07	0.000	0.000	0.000	0.000	8.08	0.000	0.000	0.000	0.000

(continued)

TABLE 4.24 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	21.89	-0.001	0.006	0.005	-0.001	21.80	-0.001	0.004	0.005	-0.001
	25%	26.10	-0.001	0.008	0.003	-0.002	25.94	-0.001	0.005	0.003	-0.001
	50%	24.46	-0.001	0.007	0.001	-0.002	24.30	-0.001	0.005	0.001	-0.001
	75%	17.78	-0.001	0.004	0.000	-0.001	17.70	0.000	0.003	0.000	-0.001
0.2											
	0%	17.10	0.000	0.004	0.003	-0.001	17.06	0.000	0.003	0.003	-0.001
	25%	23.60	-0.001	0.007	0.002	-0.002	23.52	-0.001	0.005	0.002	-0.001
	50%	23.29	-0.001	0.007	0.001	-0.002	23.21	-0.001	0.005	0.001	-0.001
	75%	17.38	0.000	0.004	0.000	-0.001	17.34	0.000	0.003	0.000	-0.001
0.5											
	0%	11.87	0.000	0.002	0.001	0.000	11.87	0.000	0.001	0.001	0.000
	25%	21.01	-0.001	0.006	0.001	-0.001	20.97	-0.001	0.004	0.001	-0.001
	50%	22.16	-0.001	0.006	0.000	-0.001	22.10	-0.001	0.004	0.000	-0.001
	75%	17.03	0.000	0.004	0.000	-0.001	17.00	0.000	0.003	0.000	0.000

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.25 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 160 and CS = 10).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	131.14	-0.007	0.032	0.075	-0.014	130.76	-0.007	0.023	0.075	-0.007
	25%	176.29	-0.009	0.038	0.051	-0.019	175.58	-0.009	0.027	0.051	-0.010
	50%	174.97	-0.009	0.038	0.030	-0.018	173.33	-0.009	0.027	0.030	-0.010
	75%	116.36	-0.006	0.029	0.004	-0.012	114.83	-0.006	0.021	0.004	-0.006
	100%	7.93	0.000	0.000	0.000	0.000	7.93	0.000	0.000	0.000	0.000
0.2											
	0%	98.58	-0.005	0.027	0.047	-0.010	98.34	-0.005	0.019	0.047	-0.005
	25%	165.24	-0.008	0.037	0.033	-0.017	164.74	-0.008	0.026	0.032	-0.009
	50%	173.82	-0.009	0.038	0.014	-0.018	172.88	-0.009	0.027	0.014	-0.010
	75%	111.72	-0.006	0.029	0.001	-0.011	111.14	-0.005	0.021	0.001	-0.006
	100%	7.92	0.000	0.000	0.000	0.000	7.92	0.000	0.000	0.000	0.000
0.5											
	0%	52.07	-0.002	0.017	0.009	-0.005	51.98	-0.002	0.012	0.009	-0.002
	25%	149.82	-0.007	0.035	0.007	-0.014	149.39	-0.007	0.025	0.007	-0.008
	50%	168.23	-0.008	0.037	0.003	-0.016	167.68	-0.008	0.027	0.003	-0.009
	75%	107.93	-0.005	0.028	0.002	-0.010	107.57	-0.005	0.020	0.002	-0.006
	100%	7.91	0.000	0.000	0.000	0.000	7.91	0.000	0.000	0.000	0.000

(continued)

TABLE 4.25 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	31.01	-0.001	0.007	0.005	-0.001	31.01	-0.001	0.007	0.005	-0.001
	25%	40.66	-0.002	0.009	0.003	-0.002	40.66	-0.002	0.009	0.003	-0.002
	50%	38.62	-0.002	0.009	0.001	-0.002	38.62	-0.002	0.009	0.001	-0.002
	75%	25.94	-0.001	0.006	0.000	-0.001	25.94	-0.001	0.006	0.000	-0.001
0.2											
	0%	24.17	-0.001	0.005	0.003	-0.001	24.17	-0.001	0.005	0.003	-0.001
	25%	37.10	-0.001	0.009	0.002	-0.002	37.10	-0.001	0.009	0.002	-0.002
	50%	36.89	-0.001	0.009	0.001	-0.002	36.89	-0.001	0.009	0.001	-0.002
	75%	25.26	-0.001	0.005	0.000	-0.001	25.26	-0.001	0.005	0.000	-0.001
0.5											
	0%	15.31	0.000	0.002	0.001	0.000	15.31	0.000	0.002	0.001	0.000
	25%	32.64	-0.001	0.007	0.000	-0.001	32.64	-0.001	0.007	0.000	-0.001
	50%	34.84	-0.001	0.008	0.000	-0.002	34.84	-0.001	0.008	0.000	-0.002
	75%	24.50	-0.001	0.005	0.000	-0.001	24.50	-0.001	0.005	0.000	-0.001

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI.

TABLE 4.26 Changes in GFIs in Response to Different Levels of ICCs, Proportion of Factor Loading Invariance, and between-group differences (CN = 160 and CS = 20).

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.5 between-group difference in intercepts									
0.1											
	0%	152.23	-0.004	0.024	0.077	-0.008	151.94	-0.004	0.018	0.077	-0.004
	25%	189.72	-0.005	0.028	0.052	-0.010	189.22	-0.005	0.020	0.052	-0.005
	50%	184.53	-0.005	0.027	0.028	-0.009	183.42	-0.005	0.020	0.028	-0.005
	75%	122.12	-0.003	0.021	0.003	-0.006	121.08	-0.003	0.015	0.003	-0.003
	100%	8.16	0.000	0.000	0.000	0.000	8.18	0.000	0.000	0.000	0.000
0.2											
	0%	108.43	-0.003	0.020	0.047	-0.005	108.23	-0.003	0.014	0.047	-0.003
	25%	173.73	-0.004	0.026	0.033	-0.009	173.30	-0.004	0.019	0.032	-0.005
	50%	181.33	-0.005	0.027	0.013	-0.009	180.52	-0.005	0.020	0.013	-0.005
	75%	117.98	-0.003	0.021	0.001	-0.006	117.42	-0.003	0.015	0.001	-0.003
	100%	8.18	0.000	0.000	0.000	0.000	8.19	0.000	0.000	0.000	0.000
0.5											
	0%	54.15	-0.001	0.012	0.009	-0.002	54.06	-0.001	0.009	0.009	-0.001
	25%	155.56	-0.004	0.025	0.008	-0.008	155.08	-0.004	0.018	0.008	-0.004
	50%	174.80	-0.004	0.027	0.003	-0.008	174.18	-0.004	0.019	0.003	-0.005
	75%	113.96	-0.003	0.020	0.002	-0.005	113.57	-0.003	0.015	0.002	-0.003
	100%	8.19	0.000	0.000	0.000	0.000	8.20	0.000	0.000	0.000	0.000

(continued)

TABLE 4.26 (continued)

ICC	Invariance (%)	Level-specific Indices					"Standard" Indices				
		I	II	III	IV	V	i	ii	iii	iv	v
		0.2 between-group difference in intercepts									
0.1											
	0%	36.04	-0.001	0.008	0.005	-0.002	35.94	-0.001	0.006	0.005	-0.001
	25%	44.02	-0.001	0.010	0.003	-0.002	43.86	-0.001	0.007	0.003	-0.001
	50%	40.66	-0.001	0.009	0.001	-0.002	40.51	-0.001	0.007	0.001	-0.001
	75%	27.69	-0.001	0.006	0.000	-0.001	27.59	-0.001	0.004	0.000	-0.001
0.2											
	0%	26.46	-0.001	0.006	0.003	-0.001	26.42	0.000	0.004	0.003	-0.001
	25%	39.13	-0.001	0.009	0.002	-0.002	39.05	-0.001	0.006	0.002	-0.001
	50%	38.41	-0.001	0.009	0.001	-0.002	38.33	-0.001	0.006	0.001	-0.001
	75%	26.88	-0.001	0.006	0.000	-0.001	26.83	0.000	0.004	0.000	-0.001
0.5											
	0%	15.96	0.000	0.003	0.001	0.000	15.95	0.000	0.002	0.001	0.000
	25%	33.88	-0.001	0.008	0.000	-0.001	33.83	-0.001	0.005	0.000	-0.001
	50%	36.04	-0.001	0.008	0.000	-0.001	35.99	-0.001	0.006	0.000	-0.001
	75%	26.04	-0.001	0.006	0.000	-0.001	26.00	0.000	0.004	0.000	-0.001

Note. ICC = Intra-class correlation.
CN: number of clusters.
CS: cluster size.
$_B$ denotes between-level.
I: $\Delta\chi^2(8)_B$, II: ΔCFI_B, III: $\Delta RMSEA_B$, IV: $\Delta SRMR_B$, V: ΔTLI_B.
i: $\Delta\chi^2(8)$, ii: ΔCFI, iii: $\Delta RMSEA$, iv: $\Delta SRMR_B$, v: ΔTLI

DISCUSSION

This study investigated the sensitivity of various model fit indices commonly used in SEM with respect to testing factor loading and intercept invariance in MSEM. In specific, the effects of ICC, number of cluster, cluster size, between-group difference, and proportion of invariance on changes in model fit indices had been inspected, in which proportion of invariance was a focus. In comparison to invariant conditions, noninvariance was examined at the between-level factor loadings and intercepts (i.e., testing metric invariance and scalar invariance). Noninvariance was only considered in the between-level model in this study, and therefore fit indices for the between-level model were reported accordingly. Specific fit indices for the within-level model were not reported for a simple presentation.

The results of the present study suggested that ICC and between-group difference affected the changes of model fit indices in MSEM. When testing invariance in the factor loadings, larger ICC resulted in larger changes in the fit indices. The picture was different when testing invariance at the intercept level. Smaller changes were observed when ICC was larger. One similarity between testing invariance at the factor loading and the intercept level was that when between-group difference was larger the changes in fit indices tended to be larger. For example, if the reference group had larger factor loadings than that of the focal group, the changes in model fit indices appeared more obvious. In contrast, the changes in model fit indices were trivial given small between-group difference. Number of cluster and cluster size played a role in the model fit changes. Larger number of cluster and cluster size resulted in fairly larger difference in

model fit. In summary, the sensitivity of model fit indices to lack of invariance in the between-level factor loadings and intercepts depended on the ICC levels and the magnitude of between-group difference in MSEM. Proportion of invariance was discussed next that affected changes of model fit indices.

Out of five model fit indices investigated in this study ($\Delta\chi^2$, ΔCFI, ΔRMSEA, ΔSRMR, and ΔTLI), ΔSRMR had the most noticeable and monotonic change in response to different proportions of invariance. When the degree of noninvariance increased, ΔSRMR inflated, which indicated severer problem of noninvariance. This pattern was equally apparent for testing invariance in factor loadings and intercepts. The other four indices (i.e., $\Delta\chi^2$, ΔCFI, ΔRMSEA, and ΔTLI) showed to be less sensitive to detect noninvariance. In response to lack of invariance, $\Delta\chi^2$ varied with a wider range than ΔCFI, ΔRMSEA, and ΔTLI. Herein ΔCFI had the smallest changes in reaction to different degrees of invariance. ΔRMSEA tended to be fairly larger than ΔCFI with respect to difference degree of invariance. ΔTLI was in between ΔCFI and ΔRMSEA. Though having more variability, $\Delta\chi^2$ did not change in a systematic way. That is, changes did not vary monotonically as the change in the degree of noninvariance. For example, larger degree of noninvariance didn't cause larger changes in χ^2. For example, when half of the items were noninvariant (i.e., 50% invariance), the changes were the largest, whereas the changes were smaller when invariant items changed to 75%. In this aspect, the performance of ΔCFI, ΔRMSEA, and ΔTLI were similar to $\Delta\chi^2$ in response to different proportions of invariance, regardless of examining invariance in factor loadings or intercepts. To summarize, $\Delta SRMR_B$ was the only index that systematically

reflected poor fit with increment in the degree of noninvariance in the between-level factor loadings and intercepts. One possible reason was because the SRMR was solely based on residuals as opposed to the other indices which depended on χ^2.

While comparing the changes in fit indices, it was observed that ΔSRMR in the level-specific approach had identical results to results obtained in the standard approach. As discussed previously, SRMR in the standard approach was also a model specific fit index. Generally, changes in χ^2, RMSEA, CFI, and TLI appeared larger in the level-specific approach compared to the standard approach. Larger changes were desired when comparing the constrained model to the unconstrained model, in which the changes evidenced potential noninvariant parameters across groups in multiple group analysis. This findings of the present study showed the advantage of level-specific model fit indices which were able to detect noninvariant parameters in the specific models (e.g., between-level model only). The level-specific model fit indices may provide important information that may be missed out using the standard approach.

Limitations, Future Directions and Strengths

The present study had its own strengths but the conclusions drawn from this study were limited due to the design of this study. First, a limited number of design factors was included in this study. Actually the range of factor that may influence the change in fit indices in MSEM can be very large. These included, but were not limited to: sample size, cluster size, ICC, the number of indicators per factor, the number of factors, indicator reliability, the number of groups, unbalance versus balance design (e.g., group size), between-group difference, the type of model, severity of

misspecification (e.g., measurement noninvariance), pattern of invariance, and so on. However, it was unwieldy and irrational to include all potential conditions in one simulation. The current study was restricted to a subgroup of those factors. Consequently, findings from this study should not be generalized to all settings.

Second, noninvariance was only simulated in the between-level factor loadings and intercepts. Further studies were needed to compare the sensitivity of model fit indices to the between-level model misspecification (measurement invariance in this study) with that of the within-level model misspecification. In addition, higher level of invariance (e.g., residual invariance) can be of interest. Chen (2007) studied the lack of invariance in the residual level in single-level SEM. Further work was expected to study high level of invariance in MSEM. Third, in this study, the noninvariance was simulated across groups at the cluster level. More complicated situations such as the grouping variable was a within-level variable posed more challenges. Kim, Yoon, and Kwok (2012) and Ryu (2014b) explored the grouping variable at the within level. Future research was needed to extend in such settings.

Fourth, Cheung and Rensvold (2002), Chen (2007), and Meade, Johnson, Braddy (2008) proposed cutoff values of ΔGFIs for rejecting measurement invariance in single-level analysis. The present study had not include discussion on cutoff criterion based on ΔGFIs. Future work was expected to include examination of cutoff values of ΔGFIs in MSEM. Fifth, this study only considered a simple one factor model with eight indicators and thus may not reflect realistically more complicated factor analyses wherein correlated factors and cross loadings may be of interest. The last limitation was

concerned with varying the proportions of invariance. This study considered five proportions of invariance (i.e., 0%, 25%, 50%, 75%, and 100%) and it was questioned what the trend of change would look like if more different proportions were considered (e.g., in a continuum). Future studies were needed for examining the change of fit indices (e.g., SRMR) in such scenario – whether the changes would still be monotonic.

Despite the abovementioned limitations, several strengths of the study can be noted. First, model fit in single-level SEM has been an active area in research for years and established rules have been disseminated such as Hu and Bentler (1999)'s cutoff criteria. However, it is important to determine whether findings from the single-level SEM literature can be generalized to MSEM settings. This study addressed this rising issue. Second, this study brought attention to the great complexity of testing factorial invariance in MSEM and had interesting findings to substantive researchers. Innovative methods of fit evaluation in MSEM need continue to be explored.

CHAPTER V

CONCLUSIONS

With increasing interest in MSEM, substantively and methodologically, it was expected that future research would continue to expand from the single-level SEM context to the MSEM context. This dissertation addressed two current issues in the MSEM context and discussed them in two self-contained studies. Though they were two different studies, they were related in some aspects and general conclusions can be made for these two studies. First, both studies targeted the MSEM framework and explored the guidelines which have been well discussed and established in the single-level analysis but not multilevel analysis. Second, both studies showed the advantages of a model-based approach in analyzing data of multilevel structure. The specification of multilevel models allowed for examining relations among variables at specific data level. Nevertheless, these two studies contributed to the MSEM literature in a distinctive way, which were summarized as follows.

Study 1 revisited the issue of application of sampling weights in large-scale, complex survey data. It targeted MSEM in investigating multilevel mediation. In examining multilevel mediation effects the weighted model-based approach was suggested for analyzing large-scale, complex survey data due to complex sampling designs. It brought attention to researchers using secondary data that the effect of sampling weights should be investigated and the application of sampling weights was encouraged.

Study 2 provided advice to researchers engaged in examining factorial invariance in multilevel data. It targeted MSEM in examining multilevel factorial invariance in multilevel factor models. ΔSRMR along with $\Delta\chi^2$ were recommended for detecting lack of invariance in factor loadings and intercepts in between-level models. Additionally, level-specific model fit evaluation was suggested for examining model misspecification due to measurement noninvariance in level-specific models in MSEM.

In conclusion, issues arising from using guidelines which were established in single-level SEM needed to be investigated in MSEM. Though due to the limitation of the study designs, this dissertation was a step forward in searching for guidelines in MSEM. Future research on MSEM model specification, MSEM modeling strategies, and model fit evaluation in MSEM were still needed.

REFERENCES

Arbuckle, J. L. (2006). *Amos* (Version 7.0) [Computer Program]. Chicago: SPSS.

Asparouhov, T. (2004), Weighting for unequal probability of selection in multilevel modeling. *Mplus Web Note, No.8.* Retrieved from http://statmodel.com/download/webnotes/MplusNote81.pdf

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(3), 411–434.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods, 35*(3), 439–460.

Asparouhov, T., & Muthén, B. (2012). Multiple group multilevel analysis. *Mplus Web Notes, No. 16*. Retrieved from www.statmodel.com/examples/webnotes/webnote16.pdf

Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*(2), 142–163.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*,238-246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA:

Multivariate Software.

Binder, D. A., & Roberts, G. R. (2001, January). Can informative designs be ignorable? *Newsletter of the Survey Research Methods Section, American Statistical Association, 12*(1), 4–6.

Bodovski, K., Nahum-Shani, I., & Walsh, R. (2013). School climate and students' early mathematics learning: Another search for contextual effects. *American Journal of Education, 119*(2), 209-234.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*(11), S176-S181.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology, 98*(4), 550–558.

Bureau of Labor Statistics, U.S. Department of Labor. *National Longitudinal Survey of Youth 1979 cohort, 1979-2010 (rounds 1-24)*. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH: 2012.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

Byrne, B., & Watkins, D. (2003). The Issue Of Measurement Invariance Revisited. *Journal of Cross-Cultural Psychology, 34*, 155-175.

Byrne, B. M., & van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107-132.

Cai, T. (2013). Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses. *Sociological Methodology, 43*(1), 178-219.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(3), 471-492.

Cheung, G. W., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15*(2), 167-198.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Cochran, W. G. (1963). *Sampling techniques.* (2nd ed.). New York: Wiley.

De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual-monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied Psycholinguistics, 35*(6), 1189-1211.

Georges, A., Brooks-Gunn, J., & Malone, L. M. (2012). Links between young children's behavior and achievement: The role of social class and classroom composition. *American Behavioral Scientist, 56*(7), 961-990.

Graubard, B. I., & Korn, E. L. (1996). Modeling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research, 5,* 263–281.

Gravers, S. L., Jr., & Wright, L. B. (2011). Parent involvement at school entry: A national examination of group differences and achievement. *School Psychology International, 32*(1), 35–48.

Grilli, L., & Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey methodology, 30*(1), 93–104.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education, 73*(3), 221-248.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling: A Multidisciplinary Journal, 7(4),* 534-556.

Heck, R. H. (2001). Multilevel modeling with SEM. In G. A. Marcoulides, & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 89-127). Mahwah, N.J.: Lawrence Erlbaum Associates.

Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2nd ed.). New York, NY: Routledge.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 4*, 179–188.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Hox, J. J., & Kreft, I. G. G. (1994). Multilevel analysis methods. *Sociological Methods & Research, 22*(3), 283-299.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(2), 157–174.

Hsu, H., Kwok, O., Lin, H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research, 50*(2), 197-215.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Jak, S. (2014). Testing strong factorial invariance using three-level structural equation modeling. *Frontiers in Psychology*, 5:745.

Jak, S., Oort, F. & Dolan, C. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 265-282.

Jak, S., Oort, F., & Dolan, C. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 31-39.

Jia, Y, Stokes, L., Harris, I., & Wang, Y. (2011). Performance of random effects model estimators under complex sampling designs. *Journal of Educational and Behavioral Statistics, 36*(1), 6-32.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8: User's reference guide* (2nd ed.). Chicago: Scientific Software International.

Jöreskog K. G. (1971) Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409–426.

Kalton, G. (1989). Modeling considerations: Discussion from a survey sampling perspective. In D. Kasprzyk, G. Duncan, G. Kalton, & M. Singh. (Eds.), *Panel surveys* (pp. 575–585). New York: Wiley.

Kaplan, D., & Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22*(3), 323-347.

Kaplan, D., & Ferguson, A. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(4), 305–321.

Kenny, D. A. (2008). *Measuring model fit.* Retrieved April 7, 2015, from http://davidakenny.net/cm/fit.htm

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods, 8*(2), 115–128.

Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(2), 250-267.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement, 72*(3), 469-492.

Korn. E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician, 49*(3), 291–295.

Kovacevic, M. S., & Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modeling of survey data. *Communications in Statistics, Theory and Methods, 32*, 103–121.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). *Analyzing complex survey data.* Newbury Park, CA: Sage.

Longford, N. T. (1995). *Model-based methods for analysis of data from 1990 NAEP trial state assessment.* Washington, DC: National Center for Education Statistics.

Lumley T. (2004). Analysis of Complex Survey Samples. J*ournal of Statistical Software. 9*, 1–19.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*(1), 593–614.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum Associates.

Matthews, J. S., & Kizzie, K. T. (2010). African Americans and boys: Understanding the literacy gap, tracing academic trajectories, and evaluating the role of learning-related skills. *Journal of Educational Psychology, 102*(3), 757–771.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Meredith W. (1964a). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Meredith W. (1964b). Rotation to achieve factorial invariance. *Psychometrika, 29*, 187–206.

Meredith W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Taylor & Francis Group.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Muthén, B. O. (1998-2004). *Mplus technical appendices.* Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods, 17,* 313-335.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599–620.

Muthén, B. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25,* 267–316.

National Center for Health Statistics. (2000). *Design and Estimation for the National Health Interview Survey, 1995-2004.* Vital Health Statistics, 2(130). http://www.cdc.gov/nchs/data/series/sr_02/sr02_130.pdf

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6,* 150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*, 107–124.

Pentz, M. A., & Chou, C. P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology, 62*(3), 450–462.

Peugh, J. L. (2013). Conducting three-level cross-sectional analyses. *Journal of Early Adolescence. 34*(1), 7-37.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique, 61*(2), 317–337.

Pfeffermann, D. (1996). The use of sampling weights for Survey Data Analysis. *Statistical Methods in Medical Research, 5*(1), 239–261.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: series B (statistical methodology), 60*(1), 23–40.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association, 87*(418), 383–396.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 12*, 879-891.

Preacher, K. J., & Selig, J. P. (2010, July). *Monte Carlo method for assessing multilevel mediation: An interactive tool for creating confidence intervals for indirect effects in 1-1-1 multilevel models*. Available from http://quantpsy.org/.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15,* 209-233.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: the advantages of multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal, 18*, 161–182.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805–827.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Roberts, J. K., & Fan, X. (2004). Bootstrapping within the multilevel/hierarchical linear modeling framework: A primer for use with SAS and SPLUS. *Multiple Linear Regression Viewpoints, 30*(1), 23-34.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16,* 583–601.

Ryu, E. (2014a). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, 5:81.

Ryu, E. (2014b). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *67*, 172-194.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173-180.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross national consumer research. *Journal of Consumer Research, 25*(1), 78–90.

Schulting, A. B., Malone, P. S. & Dodge, K. A. (2005). The effect of school-based kindergarten transition policies and practices on child academic outcomes. *Developmental Psychology, 41*(6), 860–871.

Selfa, L. A., Suter, N., Myers, S., Koch, S., Johnson, R. A., Zahs, D. A., et al. (1997). *1993 National Study of Postsecondary Faculty (NSOPF–93): Methodology report.* Washington, DC: National Center for Education Statistics.

Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In Skinner, C. J., Holt, D., & Smith, T.M.F (Eds.), *Analysis of complex surveys* (1st ed., pp. 59–88). Chichester, UK; New York, USA: Wiley.

Snijders, T. A. B. (2012). Survey weights. In Snijders, T. A. B., & Bosker R. (Eds.), *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed., pp. 216–245). Los Angeles; London: SAGE.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312.

Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 475–502.

Stapleton, L. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal, 13*(1), 28–58.

Stapleton, L. (2012). Evaluation of conditional weight approximations for two-level models. *Communications in Statistics-Simulation and Computation, 41*(2), 182-204.

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*, 495–506.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517–540.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–69.

Van de Vijver, F. J. (2009). Types of comparative studies in cross-cultural psychology. *Online Readings in Psychology and Culture, 2*(2). http://dx.doi.org/10.9707/2307-0919.1017

Wang, J. C., Xie, H. Y., & Fisher, J. H. (2011). *Multilevel models: Applications using SAS.* Berlin: De Gruyter.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324).Washington, DC: American Psychological Association.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4,* 10–18.

Wright, J. P., Morgan, M. A., Coyne, M. A., Beaver, K. M., & Barnes, J. C. (2014). Prior problem behavior accounts for the racial gap in school suspensions. *Journal of criminal justice, 42*, 257-266.

Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(1), 16–35.

Yoon, M., & Millsap, R. E. (2007) Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo Study, *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 435-463.

Yuan, K.-H., & Bentler, P. M. (2003). Eight test statistics for multilevel structural equation models. *Computational Statistics & Data Analysis, 44*, 89–107.

Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology, 37*, 53–82.

Zou, Y. Y. (2009). *The effects of parceling on testing group differences in second-order CFA models: A comparison between multi-group CFA and MIMIC models.* (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.